Urbanization and Growth[*†]

J. Vernon Henderson

Brown University

October 27, 2004

[*] I thank Diego Puga for very helpful comments on a preliminary draft of this chapter.

[†] Draft chapter prepared for Handbook of Economic Growth, Volume 1, P. Aghion and S. Durlauf (eds.), North Holland.

The study of urbanization and growth focuses on five related questions. First, why do cities form and why is economic activity so geographically concentrated in cities? In the USA, only 2% of the land area is covered by the urban built environment. This incredible geographic concentration is the central focus of economic geographers. Economists dating from Marshall (1890) have answered the question by saying urban agglomerations are based on technological externalities – the information spill-over benefits in input and output markets of having economic agents in close spatial proximity, where information decay over space is very rapid. In addition the new economic geography develops the idea that close spatial proximity involves pecuniary externalities – reduces the costs of intermediate and final good trade. Agglomeration benefits are specified typically as applying within industries or sets of inter-related industries; there is considerable debate empirically about their application across industries. That issue, as we will see later, is related to the second set of questions.

How do cities interact with each other, at any instant in time? What are the trade patterns across cities in final and intermediate outputs and how does that correspond to the roles of big and small cities? In what ways are cities specialized by either products or functions, and why? How do these patterns of specialization and diversification relate to city labor force compositions and human capital accumulations?

Given the role of cities at a point in time, the third set of questions asks how urban growth intersects with, or even defines national economic growth? The close connection between urban and national economic growth was recognized by Lucas (1988) and inspired by the development of endogenous growth models. To the extent endogenous growth is based on knowledge spillovers and sharing, given the role of close spatial proximity in spillovers, much of the interaction and sharing must occur at the level of individual cities. Given that, there must also be a close connection between economic development and urbanization. How are the two tied together? In addition the stochastic forces that shock production processes, invention, and technological progress must also play out in an urban form. How does that occur?

The fourth set of questions asks how governance, institutions, and public policy affect urbanization, which then in turn affects economic efficiency and growth. Apart from the long standing analysis of provision and financing of local public goods, there are three issues of interest specific to the urbanization process. First, public infrastructure investments in cities are enormous and the internal structure of cities affects not just the resources devoted to urban living such as commuting and congestion costs, but also affects production efficiency – the extent to which information and knowledge spillovers are fully realized and exploited. Second, institutions governing land markets, property rights, local government autonomy, and local financing including local public debt accumulation affect the city formation process, city sizes, and national economic growth. Finally, national government policies

concerning migration, trade policy, national investment in communications and transport infrastructure have profound impacts on the urban system, migration patterns, regional economic development and the like.

The final set of questions has to do with where cities locate and the economic geography of urbanization. In what regions do cities cluster and why are some regions so sparsely populated? What first nature forces of natural resource locations, including rivers and natural harbors, drive the location of economic activity? How do transport costs and technological change in transport costs affect the extent to which coastal versus hinterland regions are inhabited? And what is the role of second nature forces and history on location – how does the accumulation of economic activity based on historical market forces affect the current spatial patterns of economic activity?

This handbook chapter reviews evidence on all these questions and then turns to models that focus on aspects of the middle three questions – how do cities interact with each other; what is the relationship among urbanization, urban growth, national economic growth, and economic development; and what is the role of institutions and public policy in shaping urbanization? In terms of the first question on why cities form, there is a splendid handbook paper by Duranton and Puga (2004) reviewing models of the micro-foundations of agglomeration economies and another by Rosenthal and Strange (2004) reviewing empirical evidence on the subject. In terms of the where question, there is little in the way of models that look at the location patterns of individual cities. There are the core-periphery models of economic geography that analyze the allocation of economic activity within a country between a core and periphery region. We will discuss how these models may inform the where question for cities. But they are a topic unto themselves with excellent general handbook coverage in Overman, Venables, and Redding (2003) and coverage specific to regional issues in Ottaviano and Thisse (2004), with a review of empirical evidence in Head and Mayer (2004).

The first section reviews data and empirical evidence on aspects of the five questions. The second presents a simple system of cities model, which illustrates the basic organization of the urban sector and the interaction between economic and urban growth. The model serves as a platform to discuss issues of institutions and policy. In the third section, the model in Section 2 is adapted to analyze rural-urban transformation and urbanization as part of economic development; then policy issues for developing countries are analyzed.

1. Facts and Empirical Evidence

This section reviews basic facts and a body of empirical evidence on systems of cities. We start by looking at evidence based primarily on either the world as a whole or on large developed countries.

We look at the evolution of the size distribution of cities, Zipf's Law and related topics. Then we turn to what cities do – evidence on urban specialization and geographic concentration – and where they locate. Finally we turn to evidence that is more specific to the urbanization process in developing countries and issues surrounding that process.

1.1 The Size Distribution of Cities and Its Evolution

Work by Eaton and Eckstein (1997) on France and Japan and by Dobkins and Ioannides (2001) on the USA, with later work by Black and Henderson (2003) and Ioannides and Overman (2003) on the USA, establish some basic facts about urban systems and their development in France, Japan, and the USA over the last century or so. Foremost is that there is a wide relative size distribution of cities in large economies that is stable over time. Big and small cities coexist in equal proportions over long periods of time. Second, within that relative size distribution, individual cities are generally growing in population size over time; and what is considered a big versus small city in absolute size changes over time. Third, while there is entry of new cities and both rapid growth and decline of cities nearer the bottom of the urban hierarchy, at the top city size rankings are remarkably stable over time. Finally, size distributions of cities within countries, at least at the upper tail are well approximated by a Pareto distribution, with Zipf's Law applying in many cases. Establishing these facts raises a variety of issues and different methodological and technical approaches.

1.1.1 What is a City?

The empirical work in Eaton and Eckstein (1997) and subsequent work typically looks at the decade by decade development of urban systems. In doing so, there are critical choices researchers must make when assembling data. First is to define geographically what consists of the generic term "city". The usual definition is the "metro area", where large metro areas like Chicago comprise over 100 municipalities, or local political units. The idea in defining metro areas is to cover the entire local labor market and all contiguous manufacturing, service and residential activities radiating out from the core city, until activity peters out into farm land or very low density development. A second choice concerns how to accommodate changes in geographic definitions over time. One can use whatever contemporaneous definitions the country census/statistical bureau uses; however metro area definitions only start to be applied after World War II. Another approach is to take current metro area definitions and follow the same geographic areas back in time, focusing on non-agricultural activity.

A third problem concerns how to define "consistently" over time the threshold population size at which an agglomeration becomes a metro area, especially since the economic nature, population density, and spatial development of metro areas have changed so much over the last century. Some authors use an absolute cut-off point (e.g., urban population of 50,000 or more); some use a relative cut-off point (e.g., the minimum size city included in the sample should be .15 mean city size); and others

look at a set number (e.g., 50 or 100) of the largest cities. The relative cut-off point approach is attractive because it attempts to hold constant the area of the relative size distribution which is examined over time, as illustrated below. In presenting evidence on the topics to follow, whatever choices researchers make can strongly affect specific results. Nevertheless there are a variety of findings that are consistent across studies.

1.1.2 Evolution of the Size Distribution

In the research, one focus has been to study the evolution of the size distribution of cities, applying techniques utilized by Quah (1993) in examining cross-country growth patterns. Cities in each decade are divided by relative size into, say, 5-6 discrete categories, with fixed relative size cut-off points for each cell (e.g., < .22 of mean size, .22 to.47 of mean size, ... > 2.2 mean size). A first order Markov process is assumed and a transition matrix calculated. In many cases, stationarity of the matrix over decades can't be rejected, so cell transition probabilities are based on all transitions over time. If M is the transition matrix, i the average rate of entry of new cities in each decade (in a context where in practice there is no exit), Z the (stationary) distribution across cells of entrants (typically concentrated on the lowest cell), and f the steady-state distribution, then

$$f = \left[I - (1-i)\ M \right]^{-1}\ iZ \qquad (1)$$

In the data, relative size distributions are remarkably stable over time and steady-state distributions tend to be close to the most recent distributions. In the studies on the USA, Japan, and France, there is no tendency of distributions to collapse and concentrate in one cell, or for all cities to converge to mean size; nor generally is there a tendency for distributions to become bipolar. Distributions are remarkably stable. I illustrate this based on a world cities analysis (although, conceptually, distributions may better apply to countries, within which populations are relatively mobile).

Table 1 gives the size distribution of world metro areas over 100,000 population in 2000. Details on the data are available on-line.[1] Note that much of the world's population in cities over 100,000 are in small-medium size metro areas. 56% are in cities under 2 million, while only 17% are in cities over 8 million. Moreover all these cities only account for 62% of the world's urban population; the rest live in cities smaller than 100,000. So overall 73% of the world's urban population lives in cities under 2 million in population. While the popular press may focus on mega-cities, only a small part of the action is there.

Figure 1 plots the relative size distribution of the approximately 1200 metro areas worldwide over 100,000 in 1960 against the relative size distribution of the approximately 1700 metro areas over 200,000

[1] http://www.econ.brown.edu/faculty/henderson/worldcities.html.

in 2000. Relative sizes are actual sizes divided by the world average size in the corresponding year. The 100, 000 versus 200,000 cut-off points for minimum size are relative ones based on a constant minimum to mean size ratio. (Although using an absolute cut-off point in this case has little impact on the figure.) The figure plots the histogram for 20 cells on a log scale. The 1960 versus 2000 distributions for all cities worldwide (Figure 1a) and for those in developing and transition economies (Figure 1b) almost perfectly overlap. Relative size distributions are stable. Similarly performing transition analysis on world cities for 1960-1970-1980-1990-2000 and calculating the steady state distributions, starting with 5 cells and shares in each of .351, .299, .151, .100, .0991 in 1960, as we move up the urban hierarchy the steady state shares are .324, .299, .138, .122, and .117. Again, this indicates rock stability of distributions over time.

An alternative way of expressing this is to calculate spatial Gini's (Krugman, 1991b). For a spatial Gini rank all cities from smallest to largest on the x-axis and on the y-axis calculate their Lorenz curve -- the cumulative share of total sample population. The Gini is the share of the area below the 45° line, between that line and the Lorenz curve. The greater the Gini, the "less equal" the size distribution. The world Gini in 1960 versus 2000 is .59 versus .56 for developed countries, .57 versus .56 for less developed countries, and .52 versus .45 for transition economies as noted in Table 2 columns 1-4. Table 2 also lists Gini's for 1960 versus 2000 for 14 countries. Note apart from transition economies (and Nigeria), the lack of change; and note also that transition economies are distinctly "more equal". Transition economies have forestalled the growth of mega-cities through explicit and implicit (housing availability in cities) migration restrictions, as discussed in Section 3.3.1.

A second finding in examining city size distributions is that, for larger cities, over time there is little change in relative size rankings. In Japan and France, the 39-40 largest cities in 1925 and 1876, respectively, all remain in the top 50 in 1985 and 1990 respectively; and, at the top, absolute rankings are unchanged (Eaton and Eckstein, 1997). The USA displays more mobility due to substantial entry of new cities. However, while smaller cities do move up and down in rank, the biggest cities tend to remain big over time. So, for example, cities in the top decile of ranking stay in that decile indefinitely, with newer cities joining that decile as the total number of cities expands. Alternatively viewed, based on the Markov transition process, the mean first passage time for a city to move from the top to bottom cell is thousands of years (Black and Henderson, 2003). In the world cities data, as in the USA data, the probability in the transition matrix of moving out of the top cell to the next cell is very small: .038 in a decade time frame. Why do big cities stay big? A common answer is physical infrastructure (see Section 3.3.2). Large cities have huge historical capital stocks of streets, buildings, sewers, water mains and parks that are cheaply maintained and almost infinitely lived in, that give them a persistent comparative advantage over cities without that built-up stock. A second answer is modeled in Arthur (1990) and Rauch (1993) where, with localized scale externalities in production, large cities with a particular set of industries have a

comparative advantage in attracting new firms, relative to cities with a small representation of those industries. Large cities have an established scale, offering high levels of scale externalities, which smaller cities can only achieve quickly if they are able to co-ordinate mass in-migration of firms into their location, something which may be institutionally difficult to do.

1.1.3 Growth in City Numbers and Sizes

For any steady state size distribution of cities, as urbanization and growth proceed, both the absolute sizes and numbers of cities have grown historically, as a country's urban population expands through rural-urban migration and overall population growth. City sizes in the USA, Japan, and France over the past century have grown at average annual rates of 1.2 - 1.5%, depending on the country and exact time interval, rates which involve city sizes rising 3.3 - 4.5 fold every century. A small city today which is 250,000 would have been a major center in 1900.

In the world city data set, for comparable sets of countries the numbers of metro areas grew by 62% from 1960-2000 using a relative cut-off point (approximately 100,000 in 1960 versus 170,000 for this sample in 2000). Average sizes grew by about 70%. Decade by decade figures are given in Table 3. Using an absolute cut-off point of 100,000, numbers have about doubled and average sizes grown by 36% over 40 years. However we count cities, it is clear they have grown in population and numbers on an on-going basis over the decades.

The theory section will model city size growth and numbers in developed, or fully urbanized countries in Section 2 and in urbanizing economies in Section 3, as related to technological change induced by knowledge accumulation and demographic changes. There is empirical work relating city size increases to changes in knowledge levels. Glaeser, Scheinkman, and Shleifer (1995) in a cross-section city growth framework estimate that controlling for 1960 population, in the USA, cities in 1990 are 7% larger if they had a one-standard deviation higher level of median years of schooling in 1960. Black and Henderson (1999) place the issue in a panel context for 1940-1990 for the USA controlling for city fixed effects and industrial composition; and they examine the impact of percent college educated (which has enormous time variation). They find a one-standard deviation increase in the percent college educated leads to a 20% increase in city size over a decade.

1.1.4 Zipf's Law

In considering the size distribution of cities, especially in a cross-sectional context, there is a large literature on what is termed Zipf's Law (e.g., Rosen and Resnick 1980, Clark and Stabler 1991, Mills and Hamilton 1994, and Ioannides and Overman 2003). City sizes are postulated to follow a Pareto distribution, where if R is rank from smallest, r, to largest, 1, and n is size

$$R(n) = An^{-a} \tag{2}$$

given the Prob $(\tilde{n} > n) = An^{-a}$ and relative rank is $R(n)/r$, or the proportion of cities with size greater than n. Under Zipf's Law $a = 1$, or we have the rank size rule where, for every city, rank times size is a constant, A. Putting (2) in log-linear form, empirical work produces a's that vary across countries, samples, and times; but many are "close" to one. This empirical regularity has drawn considerable attention and is often used to characterize spatial inequality, using (2) as a first approximation of the true size distribution. We list sample a coefficients for 2000 for fifteen countries in Table 2, column 5. Note however while people often say that an exponent of .74 or 1.34 is "close to" one, such coefficients produce very different city size distributions, than if the coefficient is one.[2] As a declines, or the slope of the rank size line gets flatter, urban concentration is viewed as increasing: for given size changes, rank changes more slowly, or cities are "less equal". In Table 2, the a coefficients and the Gini's are in fact strongly negatively correlated, as one would expect. But we note that typically the log version of equation (2) is better approximated by a quadratic form than linear one. However one looks at it, Zipf's Law is just an approximation that does well in some circumstances and not so well in others. If one wants to compare measures of urban concentration across countries or over time, rather than compare estimated *a*'s in eq. (2), using Gini's may be more reliable, just as they are for comparisons of income distributions.

If Zipf's Law holds even approximately, why is that? In an interesting development, Gabaix (1999a, 1999b) starts to formalize the underlying stochastic components which might lead to such a relationship, building on Simon (1955). Gabaix shows that if city growth rates obey Gibrat's Law where growth rates are random draws from the same distribution,[3] so growth rates are independent of current size, Zipf's Law emerges as the limiting size distribution (as long as a lower bound on how far cities can deteriorate in size is imposed). Growth is scale invariant, so the final distribution is; and we have a power law with exponent 1. Gabaix sketches an illustrative model, based on on-going natural amenity shocks facing cities of any size, which leads to Zipf's Law for the size distribution of cities. More comprehensive formulations in Duranton (2004) and Rossi-Hansberg and Wright (2004) are discussed in the theory section.

While Gibrat's Law is a neat underlying stochastic process, does it hold up empirically? Black and Henderson (2003) test whether in the relationship, $\ln n_{it} - \ln n_{it-1} = a + \delta t + \alpha \ln n_{it-1} + \varepsilon_{it}$, $\alpha = 0$ as hypothesized under the Law. The Law requires ε_{it} to be i.i.d., so simple OLS suffices. Black and Henderson find $\alpha < 0$ under a variety of circumstance and sub-samples, under appropriate statistical criteria, which rejects Gibrat's Law. Ioannides and Overman (2003) examine the issue more thoroughly in a non-parametric fashion, characterizing the mean and variance of the distribution from which growth

[2] I don't report standard errors since OLS estimates of standard errors are biased downwards. See Gabaix and Ioannides (2004).

[3] Actually the requirement is that they face the same mean and variance in the drawing.

rates are drawn. The mean and variance of growth rates do seem to vary with city size but bootstrapped confidence intervals are fairly wide generally, allowing for the possibility of (almost) equal means.

1.2 Geographic Concentration and Urban Specialization

Geographic concentration refers to the extent to which an industry k is concentrated at a particular location or, more generally concentrated at a few versus many locations nationally. A common measure of concentration of industry k at location i is $l_{ik} = X_{ik} / \sum_i X_{ik}$, where X_{ik} is location i 's employment or output of industry k. Thus l_{ik} is location i 's share of, say, national employment in industry k. In contrast to geographic concentration, specialization refers to how much of a location's total employment is found in industry k, or $s_{ik} = X_{ik} / \sum_k X_{ik}$. As Overman, Redding and Venables (2001) demonstrate, if we normalize l_{ik} by location i 's share of national employment $(s_{ik} \equiv \sum_k X_{ik} / \sum_k \sum_i X_{ik})$ and s_{ik} by industry k 's share of national employment $(s_k \equiv \sum_i X_{ik} / \sum_k \sum_i X_{ik})$ we get the same measure -- a location quotient, or

$$q_{ik} = X_{ik} \frac{(\sum_k \sum_i X_{ik})}{\sum_k X_{ik} \sum_i X_{ik}} \qquad (3)$$

The distribution of q_{ik} across industries, k, compared over time for a city would tell us about how city i 's specialization patterns are changing over time. And the distribution of q_{ik} across locations, i, over time would tell us whether industry k is becoming more or less concentrated over time at different locations. In a practical applications looking at many industries and cities over time or across countries, the issue concerns how to produce summary measures to describe either how overall concentration varies across industries or how one city's specialization compares with another's. Another issue concerns how to factor in the different forces that cause specialization or concentration phenomena. The literature uses a variety of approaches. We start by looking at urban specialization.

1.2.1 Urban Specialization

Evidence on countries such as Brazil, U.S.A., Korea, and India (Henderson 1988, and Lee 1997) indicate that cities are relatively specialized. The traditional urban specialization literature going back to Bergsman, Greenston and Healy (1972) uses cluster analysis to group cities into categories based on similarity of production patterns -- correlations (or minimum distances) in the shares of different industries in local employment, s_{ik}. Cluster analysis is an "art form" in the sense that there is no optimal set of clusters, and it is up to the researcher to define how fine or how broad the clusters should be and there are a variety of clustering algorithms.

Using 1990 data for the U.S.A., Black and Henderson (2003) group 317 metro areas into 55 clusters, "defining" 55 city types based on patterns of specialization for 80 2-digit industries. They define textile, primary metals, machinery, electronics, oil and gas, transport equipment, health services, insurance, entertainment, diversified market center, and so on type cities, where anywhere from 5-33% of local employment is typically found in just one industry. They show that production patterns across the types are statistically different and that average cities and educational levels by type differ significantly across many of the types. Specialization especially among smaller cities tends to be absolute. At a 3-digit level many cities have absolutely zero employment in a variety of categories. So in the 1992 Census of Manufactures for major industries like computers, electronic components, aircraft, instruments, metal working machinery, special machinery, construction machinery, and refrigeration machinery and equipment, respectively, of 317 metro areas 40%, 17%, 42%, 15%, 77%, 15%, 14% and 24% have absolutely zero employment in these industries.

Kim (1995) in looking at the USA examines how patterns of specialization have changed over time, by comparing for pairs (i, j) of locations $\sum_k |s_{ik} - s_{jk}|$ and by estimating spatial Gini's for industry concentration. He finds that states are substantially less specialized in 1987 than in 1860, but that localization, or concentration has increased over time. For Korea, as part of the deconcentration process noted earlier, Henderson, Lee, and Lee (2001) find that from 1983 to 1993, city specialization as measured by a normalized Hirschman-Herfindahl index

$$g_j = \sum_k (s_{jk} - s_j)^2 \qquad (4)$$

rises in manufacturing, while a provincial level index declines. Cities become more specialized and provinces less so. Clearly the geographic unit of analysis matters, as do the concepts. City specialization as envisioned in the models presented below is consistent with regional diversity, when large regions are composed of many cities of different types.

Henderson (1997) for the USA and Lee (1997) for Korea show that the g_j index of specialization in manufacturing declines with metro area size. Smaller cities are much more specialized than larger cities in their manufacturing production. More generally, Kolko (1999) demonstrates that larger cities are more service oriented and smaller ones more manufacturing oriented. For six size categories (over 2.5 million, 1 - 2.5 million, ... < .25 million, non-metro counties) Kolko shows that the ratio of manufacturing to business service activity rises from .68 to 2.7 as size declines, where manufacturing and business services account for 35% of local private employment. The other 65% of local employment is in "non-traded"

activity whose shares don't vary across cities – consumer services, retail, wholesale, construction, and utilities.

1.2.2 Geographic Concentration

What about concentration of industry -- the extent to which a particular industry is found in a few versus many locations? In an extremely important paper, Ellison and Glaeser (1999) model the problem using USA data, to determine the extent of clustering of plants within an industry due to either industry-specific natural advantages (e.g., access to raw materials) or spillovers among plants. Plants locate across space so as to maximize profits and profits depend on area specific natural advantage, spillovers, and an i.i.d. drawing from Weibull distribution. The idea is to explain the joint importance of spillovers and natural advantage in geographic concentration.

Geographic concentration for industry j is $G_j = \sum_i (s_{ji} - x_i)^2$, where s_{ji} is the share of industry j in employment in location i and x_i is location i's share in total national employment (to standardize for location size). Where $0 \leq \gamma^{na} \leq 1$ represents the importance of natural advantage (where the variance in relative profitability of a location is proportional to γ^{na}) and γ^S represents the fraction of pairs of firms in an industry between which a spillover exists, under their assumptions, Ellison and Glaeser show that

$$E\left[G_j\right] = (1 - \sum_i x_i^2)\,(\gamma_j + (1-\gamma_j)\,H_j)$$
$$\gamma_j \equiv \gamma_j^{na} + \gamma_j^s - \gamma_j^s \gamma_j^{na} \qquad (5)$$

where H_j is the standard Hirschman-Herfindahl index of plant industrial concentration in industry j. So $E\left[G_j\right]$ adjusts γ_j for variations in location size $(1 - \sum x_i^2)$ and industry concentration H. Using (5) and estimates of G_j, H_j, and $(1 - \sum x_i^2)$, the empirical part of their paper calculates γ_j for all 3- or 4-digit manufacturing industries across states and countries. They show for 4-digit industries that $G > (1 - \sum x_i^2)H$ in 446 of 459 industries, where $G \leq (1 - \sum x^2)H$ only if $\gamma \leq 0$. That is, almost all industries display some degree of spatial concentration due to either natural advantage or spillovers. Second they argue that 25% of industries are highly concentrated $(\gamma > .05)$ and 43% are not highly concentrated $(\gamma < .02)$. In a later article, Ellison and Glaeser (1999) argue that, based on econometric results relating location choices to natural advantage measures, 10-20% of γ in eq. (5) is accounted for by natural advantage. The rest is due to intra-industry spillovers, a rather critical finding in urban analysis indicating the importance of understanding the nature of scale externalities.

In an important working paper, Duranton and Overman (2004), look at geographic concentration using British data. Rather than model the underlying stochastic process of industrial location under specific assumptions to yield a specific index, Duranton and Overman take a non-parametric approach, where they also focus on how to test statistically whether industries are significantly concentrated. They calculate the distribution of all pair-wise distances between plants in an industry. Distributions shifted to the left have a greater concentration of short pair-wise distances and are more spatially concentrated. The authors have the advantage of knowing "exact" plant locations (basically within a city block or so), rather than having to rely on, say, county locations, which in the US can cover vast distances. They develop a framework to test observed industry distributions against the "counterfactual" of what distributions would look like if firms choose locations randomly, given (a) the set of locations in the UK for industrial plants is limited, (b) bilateral distances between all possible points are not independent, and (c) industry sizes or numbers of plants differ. The framework involves repeated sampling for an industry without replacement from the set of national industrial sites with the sample size equal to industry size. Following that procedure, they construct 95% confidence intervals to test if observed distributions depart from randomness.

Compared to Ellison-Glaeser, in practical applications their approach captures a nuanced aspect of spatial clustering. For relatively concentrated industries, the Ellison and Glaeser index is typically dominated by the county with the highest share (given squared shares in the index), telling us the extent to which an industry is concentrated in just one place. The Duranton-Overman approach tells us more generally about spatial clustering over the whole country. So in Ellison and Glaeser, an industry which has a high concentration in one county but is otherwise very dispersed across the 3000 USA counties may look more concentrated than an industry which is concentrated in, say 3-4 nearby counties, with little representation elsewhere. But the latter would be well represented in Duranton and Overman.

1.2.3 Geography

A variety of recent studies have examined the role of geography, primarily natural features, in the spatial configuration of production and growth of cities. Rappaport and Sacks (2001) herald the role of coastline location in the U.S.A., as a factor promoting city growth. In a related study, Beeson, DeJong and Troeskan (2001) look at USA counties from 1840-1990. They show that iron deposits, other mineral deposits, river location, ocean location, river confluence, heating degree days, cooling degree days, mountain location, and precipitation all affect the base 1840 county population significantly. However for 1840-1990 *growth* in county population, only ocean location, mountain location, precipitation, and river confluence matter, controlling for 1840 population. That is, first nature items strongly affected 1840 and hence indirectly 1990 populations; but growth from 1840-1990 is independent of many first nature influences. Ocean location as Sacks' suggests has persistent growth effects.

Both these studies ignore the geography of markets and the role of neighbors in influencing city evolution. Dobkins and Ioannides (2001) show that growth of neighboring cities influences own city growth and cities with neighbors are generally larger than isolated cities. Black and Henderson (2003) put neighbor and geographic effects together. They calculate normalized market potential variables (sum of distance discounted populations of all other counties in each decade, normalized across decades). They find climate and coast affect relative city growth rates; but market potential has big effects as well, although they are non-linear. Bigger markets provide more customers, but also more competition, so marginal market potential effects diminish as market potential increases. High market potential helps explain why North-East cities in the USA maintain reasonable growth, given for historical reasons, they are in the most densely populated area, despite the hypothesized natural advantages of the West.

1.3 Urbanization in Developing Countries

Urbanization, or the shift of population from rural to urban environments, is typically a transitory process, albeit one that is socially and culturally traumatic. As a country develops, it moves from labor-intensive agricultural production to labor being increasingly employed in industry and services. The latter are not land-intensive and are located in cities because of agglomeration economies. Thus urbanization moves populations from traditional rural environments with informal political and economic institutions to the relative anonymity and more formal institutions of urban settings. That in itself requires institutional development within a country. It spatially separates families, particularly by generation, as the young migrate to cities and the old stay behind.

Urbanization is a spatial transition process. By upper middle income ranges, countries become "fully" urbanized, in the sense that the percent urbanized levels out at 60-90% of the national population living in cities. The actual percent urbanized with full urbanization varies with geography, the role of modern agriculture in the economy, and national definitions of urban. This idea of a transitory phenomenon is illustrated Figure 2, comparing different regions of the world in 1960 versus 1995. While urbanization increased in all regions of the world over those 35 years, among developed countries there is little change since 1975. By 1995 Soviet bloc and Latin American countries had almost converged to developed country urbanization levels. Only sub-Saharan African and Asian countries still face substantial urbanization in the future. Although urbanization is transitory, given the total spatial transformation and accompanying institutional and social transformation involved, as a policy issue, urbanization is very important to developing countries. Here we review some basic facts and issues about the process.

1.3.1 Issues Concerning Overall Urbanization

As noted above, urbanization is the consequence of changes in national output composition from rural agriculture to urbanized modern manufacturing and service production. As such, Renaud (1981) makes the basic point that government policies bias, or influence urbanization through their effect on national sectoral composition. So policies affecting the terms of trade between agriculture and modern industry or between traditional small town industries (textiles, food processing) and high tech large city industries affect the rural-urban or small-big city allocation of population. Such policies include tariffs, and price controls and subsidies. The idea that government policies affect urbanization primarily through their effect on sector composition is a key point of empirical studies of urbanization by Fay and Opal (1999) and Davis and Henderson (2003). These studies show that, indeed, urbanization which occurs in the early and middle stages of development is determined largely by changes in national economic sector composition and in technology. Government policies tend to affect urbanization only indirectly through their effect on sector composition.

Urbanization promotes benefits from agglomeration such as localized information and knowledge spillovers and thus efficient urbanization promotes economic growth. Writers such as Gallup, Sacks and Mellinger (1999) go further to suggest that urbanization may "cause' economic growth, rather than just emerge as part of the growth process. The limited evidence so far suggests urbanization doesn't cause growth per se. Henderson (2003) finds no econometric evidence linking the extent of urbanization to either economic or productivity growth or levels. That is, if a country were to enact policies to encourage urbanization per se, typically that wouldn't improve growth.

Finally on urbanization, there is an informal notion (Mills and Becker, 1986, and World Bank, 2000) that the transitory urbanization process follows the same stages as population growth (the "demographic" transition between falling death rates and falling fertility rates) – an S-shaped relationship where urban population growth is slow at low levels of development, then there is a period of rapid acceleration in intermediate stages, followed by a slowing of growth. However the data suggest otherwise at least over the last 35 years. Figure 3 illustrates after parceling out the effect of national population, or country size, based on pooled country data every 5 years from 1965-1995. In Figure 3 the log of national urban population is an increasing concave function of the log of income per capita, indicating the *growth* rate of urban population is a concave increasing function of income levels (Davis and Henderson, 2003).

1.3.2 The Form of Urbanization: The Degree of Spatial Concentration

In 1965, Williamson published an innovative paper based on cross-section analysis of 24 countries in which he argued that national economic development is characterized by an initial phase of internal regional divergence, followed by a phase of later convergence. That is, a few regions initially experience accelerated growth relative to other (peripheral) regions, but later the peripheral regions start

to catch up. Barro and Sala-i-Martin (1991 and 1992) present extensive evidence on this for the USA, Western Europe, and Japan, by examining the evolution of inter-regional differences in per capita incomes. While inter-regional out-migration from poorer regions plays a role in catch-up, it may not be critical. For Japan, the authors argue that later convergence of backward regions occurred mostly through improved productivity in backward regions.

The urban version of this divergence-convergence phenomenon looks at urban primacy. Following Ades and Glaeser (1995), conceptually the urban world is collapsed into two regions -- the primate city versus the rest of the country, or at least the urban portion thereof. The basic question concerns to what extent urbanization is concentrated, or confined to one (or a few) major metro areas, relative to being spread more evenly across a variety of cities. Primacy is commonly measured by the ratio of the population of the largest metro area to all urban population in the country (Ades and Glaeser 1995, Junius 1999, and Davis and Henderson, 2003). A more comprehensive measure might use either a spatial Gini or a Hirschman-Herfindal index [HHI] from the industrial organization literature.

Corresponding to Williamson's hypothesis, these papers find an inverted U-shape relationship, where relative urban concentration first increases, peaks, and then declines with economic development. Despite different concentration measures and methods, Wheaton and Shishido (1981) examining a HHI using cross-section non-linear OLS and Davis and Henderson (2003) examining primacy using panel data methods and IV estimation find that concentration rises, peaks in the $2000-4000 range (1985 PPP dollars), and then declines. As Figure 4 illustrates, without conditioning on other variables affecting primacy, the inverted $U-$ relationship of primacy against income is noisy and only apparent in the raw data in earlier time periods (cf. 1965-75 in part (a) with 1985-95 in part (b)).

Lee (1997) explores the relationship between changes in urban concentration and industrial transformation for Korea. The idea is that manufacturing is also first very concentrated in primate cities at early stages of development and then decentralizes to such an extent that at the other end of economic development it is relatively more concentrated in rural areas, as in the USA today, as noted earlier. Seoul's urban primacy peaked around 1970 and while Seoul's absolute population has continued to grow, its share has declined steadily. At the urban primacy peak in 1970, Seoul had a dominant share of national manufacturing although the other major metro areas, Pusan and Taegu, also had large shares. During the next 10-15 years, manufacturing suburbanized from Seoul to satellite cities in the rest of Kyonggi province (its immediate hinterland), as well as to satellite cities surrounding Pusan and Taegu. Such suburbanization of manufacturing has been documented also for Thailand (Lee, 1988), Colombia (Lee, 1989), and Indonesia (Henderson, Kuncoro and Nasution, 1996). But the key development following the early 1980's in Korea is the spread of manufacturing from the three major metro areas (Seoul, Pusan, and Taegu) and their satellites to rural areas and other cities. The share of rural areas and other cities in

manufacturing rose from 26% in 1983 to 42% in 1992, in a time period when national manufacturing employment is fairly stagnant and rural areas and other cities actually continue to experience modest absolute population losses. That is, manufacturing deconcentrated both relatively and absolutely to hinterland regions. This deconcentration coincided with economic liberalization, enormous and widespread investment in inter-regional transport and infrastructure investment, and fiscal decentralization (Henderson, Lee, and Lee, 2001) and is consistent with core-periphery reversal in the new economic geography literature discussed later.

Given the urban primacy relationships, the immediate issue is the "so-what" question. How is urban concentration important to growth? For example, is there an optimal degree of urban primacy with each level of development, where significant deviations from this level detract from growth? Conceptually there should be an optimal degree of primacy, where optimal primacy involves a trade-off of the benefits of increasing primacy-- enhanced local scale economies contributing to productivity growth-- against the costs -- more resources diverted away from productive and innovative activities to shoring up the quality of life in congested primate cities. In the first econometric examination of this so-what question, Henderson (2003), using panel data and IV estimation for 1960-1990, finds that there is an optimal degree of primacy at each level of development that declines as development proceeds. Optimal primacy is the level that maximizes national productivity growth. Initial high relative agglomeration is important at low levels of development when countries have low knowledge accumulation, are importing technology, and have limited capital to invest in widespread hinterland development. However the desirability of high relative agglomeration declines with development. Error bands about optimal primacy numbers are quite tight. Second, large deviations from optimal primacy strongly affect productivity growth. An 33% increase or decrease in primacy from a typical best level of .3 reduces productivity growth by 3 percentage points over five years, a big effect. There is some tendency internationally to excessive primacy, with the usual suspects such as Argentina, Chile, Peru, Thailand, Mexico, and Algeria having extremely high primacy.

Why would countries significantly deviate from desired levels of concentration? There is a considerable literature on how government policies and institutions foster excessive concentration. In Ades and Glaeser (1995), the basic idea is that national policy makers favor the national capital (or other seat of political elites such as São Paulo in Brazil) for reasons of personal gain. For example, direct restraints on trade for hinterland cities such as inability to access capital markets or to get export or import licenses favor firms in the national capital. Policy makers and bureaucrats may gain as shareholders in such firms, or they may gain rents from those seeking licenses or other exemptions to trade restraints (see Henderson and Kuncoro, 1996, on Indonesia). Indirect trade protection for the primate city can also involve under-investment in hinterland transport and communications infrastructure.

Whether as true beliefs or as a justification to cover rent-seeking behavior, policy makers in different countries often articulate a view that large cities are more productive and thus should be the site for government-owned heavy industry (e.g., São Paulo or, Beijing-Tianjin historically). Later we will point out that it may be that output per worker in heavy industries is higher in the productive external environment of large metro areas. It just isn't high enough to cover the higher opportunity costs of land and labor in those cities, which is one reason why those state-owned heavy industries lose money in such cities.

Favoritism of a primate city creates a non-level playing field in competition across cities. The favored city draws in migrants and firms from hinterland areas, creating an extremely congested high cost-of-living metro area. Local city planners can try to resist the migration response to primate city favoritism by, for example, refusing to provide legal housing development for immigrants or to provide basic public services in immigrant neighborhoods. Hence the development of squatter settlements, bustees, kampongs and so on. But still, favored cities tend to draw in enormous populations.

Is there econometric evidence indicating that politics plays a role in increasing sizes of primate cities? Ades and Glaeser (1995) based on cross-section analyses find that, if the primate city in a country is the national capital, it is 45% larger. If the country is a dictatorship, or at the extreme of non-democracy, the primate city is 40-45% larger. The idea is that representative democracy gives a political voice to the hinterland regions limiting the ability of the capital city to favor itself. Apart from representative democracy, fiscal decentralization helps to level the playing field across cities, by giving political autonomy for hinterland cities to compete with the primate city.

Davis and Henderson (2003) explore these ideas further, examining in a panel context the impact upon primacy of democratization and fiscal decentralization from 1960-1995. Using a panel approach with IV estimation, they find smaller effects than Ades and Glaeser, but still highly significant ones. Examining both democratization and fiscal decentralization together, they find moving from the extreme of least to most democratic form of government reduces primacy by 8% and from the extreme of most to least centralized government reduces primacy by 5%. Primate cities which are national capitals are 20% larger and primate cities in planned economies with migration restrictions are 18% smaller. Finally they find transport infrastructure investment in hinterlands which opens up international markets to hinterland cities reduces primacy, as the core-periphery models of the new economic geography tend to predict. A one-standard deviation increase in either roads per sq. kilometer of national land area or navigable inland waterways per sq. kilometer each reduce primacy by 10%.

2. Cities and Growth

To establish the links between cities, growth, urbanization, urban concentration and policy, we look at models in which cities are a defined unit, endogenous in number and size. These are systems of cities models which date to Henderson (1974), with a variety of substantial contributors to further development (Hochman 1977, Kanemoto 1980, Henderson and Ioannides 1981, Abdel-Rahman and Fujita 1990, Helsley and Strange 1990, Duranton and Puga 2000, and Rossi-Hansberg and Wright 2004, to name a few). Here I outline the model in Black and Henderson (1999a) which is an endogenous growth model of cities, examining the growth-urban connection. The analysis is broken into several parts. The first reviews the traditional static model, focused on city formation and the determination of the sizes, numbers, and industrial composition of cities in an economy at a point in time. A thorough review of static models is in Abdel-Rahman and Anas (2004), so our treatment focuses on what we need to analyze growth, and later urbanization and development. We then turn to the growth part, focusing on steady-state growth and a variety of extensions covering stochastic processes and analysis of functional specialization. Section 3 turns to rural-urban transformation, or urbanization under economic development. That section also discusses issues of city debt finance and land market institutions.

2.1 The Systems of Cities at a Point in Time

Consider a large economy composed of two types of cities, where there are many cities of each type and each type is specialized in the production of a specific type of traded good. We will show why (when) there is specialization momentarily; the generalization to many types of goods and cities is straightforward. To simplify the growth story, each firm is composed of a single worker. In a city type 1, in any period, the output of firm i in a type 1 city is

$$X_{1i} = D_1 \ (n_1^{\delta_1} h_1^{\psi_1}) \ h_{1i}^{\phi_1}, \ 0 < \delta_1 < \frac{1}{2} \qquad (6)$$

h_{1i} is the human capital of the worker and is his input in the production process. A worker-firm is subject to two local externalities. First is own industry localization economies, the level of which depends on the total number of worker-firms, n_1, in this representative type 1 city. There is a large literature on micro-foundations of localization economies, with an excellent analysis and review in Duranton and Puga (2004). While the concepts are discussed in Marshall (1890), the formal literature dates to Fujita and Ogawa (1982) who model micro-foundations as exogenous information spillovers that enhance productivity but decay with spatial distance between plants. Such spillovers can be made endogenous (Kim, 1988) with the volume of costly "contacts" being a firm choice variable. But the modern literature

on micro-foundations as reviewed in Duranton and Puga moves on to try to model why contacts matter, rather than just assuming they matter.

In this section spatial decay is all or nothing – no decay within the city's; 100% across cities. As such in (6), n_1 could represent the total volume of local spillover communications, where δ_1 is the elasticity of firm output with respect to n_1. The restriction $\delta_1 < 1/2$ which limits the degree of scale economies ensures a unique solution in an economy composed of many type 1 cities. Without the restriction, all X_1 production crowds into just one city. Note the production process ignores land, collapsing the central business district [CBD] to a point. There is a recent literature building upon Fujita and Ogawa where firm density is endogenous in a spatial CDB with information spillover decay. There market equilibrium density is non-optimal because firms in making location decisions don't recognize that choices leading to greater densities would enhance information spillovers (Lucas and Rossi-Hansberg, 2002 and Rossi-Hansberg, 2004). The issue of central city design and zoned density is an important one in the design of cities in developing countries. But it is beyond the scope of this review.

The second externality in (6) derives from h_1, the average level of human capital in the city, which represents local knowledge spillovers. $h_1^{\psi_1}$ could be thought of as the richness of information spillovers $n_1^{\delta_1}$, so that knowledge enhances (multiplies) local information spillovers, or gives better information. Alternatively it could just represent the level of local technology, which increases as average education increases locally.

Given this simple formulation the wage of worker i in city type 1 is

$$W_{1i} = X_{1i} \tag{7}$$

In an economy of identical individual workers in type 1 cities, individuals will all have the same human capital level (either exogenously in a static context, or endogenously in a growth context). Thus total city output is

$$X_1 = D_1 h_1^{\phi_1 + \psi_1} n_1^{1+\delta_1} \tag{8}$$

2.1.1 Equilibrium City Sizes

Equations (6) and (8) embody the scale benefits of increases in local employment, where output per worker is an increasing function of local own industry scale. Determinant city sizes arise because of scale diseconomies in city living, including per capita infrastructure costs, pollution, accidents, crime, and

commuting costs. In Henderson (1974) those are captured in a general cost of housing function, but most urban models consider an explicit internal spatial structure of cities. As noted all production occurs at a point, the CBD. Surrounding the CBD in equilibrium in local land markets is a circle of residents each on a lot of unit size. People commute back and forth at a constant cost per unit (return) distance of τ. That cost can be from working time, or here an out-of-pocket cost paid in units of X_1. Equilibrium in the land market is characterized by a linear rent gradient, declining from the center to zero at the city edge where rents (in agriculture) are normalized to zero. Standard analysis dating to Mohring (1961) gives us expressions for total city commuting and rents, in terms of city population where[4]

$$\text{total commuting costs} = bn_1^{3/2} \tag{9}$$

$$\text{total land rents} = 1/2\ bn_1^{3/2} \tag{10}$$

$$\text{b} \equiv 2/3\ \pi^{-\frac{1}{2}}\tau.$$

Equation (9) represents the key resource costs, where marginal commuting costs are increasing in city population. Rents are income to, potentially, a city developer or to rentiers.

How do cities form and how are sizes determined? We start with a specific mechanism and discuss how it generalizes below, and what happens if such a mechanism isn't present. There is an unexhausted supply of identical city sites in the economy, each owned by a land developer in a nationally competitive urban land development market. A developer for an occupied city collects local land rents, specifies city population (but there is free migration in equilibrium), and offers any inducements to firms or people to locate in that city, in competition with other cities. Population is freely mobile.

The land developer maximizes

[4] An equilibrium in residential markets requires all residents (living on equalize size lots) to spend the same amount on rent, $R(u)$, plus commuting costs, τu, for any distance u from the CBD. Any consumer then has the same amount left over to invest or spend on all other goods. At the city edge at a radius of u, rent plus commuting costs are τu_1 since $R(u_1) = 0$; elsewhere they are $R(u) + \tau u$. Equating those at the city edge with those amounts elsewhere yields the rent gradient $R(u) = \tau(u_1 - u)$. From this, we calculate total rents in the city to be $\int_0^{u_1} 2\pi u R(u) du$ (given lot sizes of one so that each "ring" $2\pi u du$ contains that many residents) or $1/3\pi\tau u_1$. Total commuting costs are $\int_0^{u_1} 2\pi u\ (\tau u)\ du = 2/3\pi r u_1^3$. Given a city population of n and lot sizes of one, $n_1 = \tau u_1^2$ or $u_1 = \pi^{-\frac{1}{2}} n^{\frac{1}{2}}$. Substitution gives us eqs. (9) and (10).

$$\max_{n_1,\ T_1} \ profit_1 = 1/2\ bn_1^{3/2} - T_1 n_1$$
$$\text{subject to } W_1 + T_1 - 3/2\ bn_1^{\frac{1}{2}} = I_1 \tag{11}$$

where T_1 is the per firm subsidy (e.g., in practice, in a model with local public goods, a tax exemption). I_1 is the real income per worker available in equilibrium in national labor markets under free mobility, which a single developer takes as given. In the constraint, I_1 equals wages in (7), plus the subsidy, less per worker rents plus commuting costs paid from (9) and (10). Maximizing with respect to T_1 and n_1 and imposing perfect competition in national land markets so $profit_1 = 0$ ex post, yields

$$T_1 = 1/2\ bn_1^{\frac{1}{2}} \tag{12}$$

$$n^* = (\delta_1 2b^{-1} D_1)^{2/(1-2\delta_1)} h_1^{2\varepsilon_1} \tag{13}$$

$$\varepsilon_1 \equiv \frac{\phi_1 + \psi_1}{1 - 2\delta_1} \tag{14}$$

This solution has a variety of properties heralded in the urban literature. First it reflects the Henry George Theorem (Flatters, Henderson, and Mieszkowski 1974, Stiglitz 1977), where the transfer per worker/firm exactly equals the gap $(\delta_1 W_1)$ between social and private marginal of labor to the city, and that subsidy which prices externalities is exactly financed out of collected land rents at efficient city size. That is, total land rents cover the cost of subsidies needed to price externalities, as well as the costs of local public goods in a model where good goods are added in. Second the efficient size in (13) is the point where real income, I_1, peaks, as an inverted $U-$shape function of city size, as we will illustrate later in Figure 5. If $\delta_1 < 1/2$, we can show that I_1 is a single-peaked function of n_1, so n_1^* is the unique efficient size. If $\delta_1 > 1/2$, in essence there will only be one type 1 city in the economy, because net scale economies are unbounded. Given n_1^* is the size where I_1 peaks, n_1^* is a free mobility equilibrium -- a worker moving to another city would lower real income in that city and be worse off. Finally city size is increasing in technology improvements: τ declining, δ_1 rising, D_1 rising, or local knowledge accumulation (h_1) rising.

By substituting in the constraint in (11), we can define relationships among real income, wages, and human capital. Substituting in first for T_1 and then n_1 we get

$$I_1 = W_1 - bn^{1/2} = (1-2\delta)\ W_1 = Q_1 h_1^{\varepsilon_1} \tag{15}$$

where Q_1 is a parameter cluster. Note real income is wages deflated by urban living costs; and that real income rises with human capital.

Institutions and City Size. I have specified the equilibrium in national land markets, given competitive developers. Helsley and Strange (1990) put this in proper context, specifying the city development game, determining how many cities will form and what their sizes (n^*) will be. Henderson and Becker (2001) show that the resulting solutions (with multiple factors of production) are (1) Pareto efficient, (2) the only coalition proof equilibria in the economy, (3) unique under appropriate parameters, and (4) free mobility ones where the developer specified populations are self-enforcing. They also show that, under appropriate conditions, such outcomes arise (1) in an economy with no developers but with city governments, where city governments can exclude residents ("no-growth" restrictions) to maximize the welfare of the representative local voter; and (2) in a growing economy where developers form new cities and old cities are governed by passive local governments. Note for developing countries the key ingredients: either national land markets must be competitive with developers free to form new cities or atomistic settlements can arise freely and local autonomous governments can limit their populations as they grow (as well as provide infrastructure once that is accounted for—see section 3.3.2).

Absent such institutions, cities only form through "self-organization". In the model here with perfect mobility of resources, the result is potentially enormously oversized cities (Henderson 1974, Henderson and Becker 2001). Nash equilibrium city size in atomistic worker migration decisions lies between efficient size, n^*, and a limit size to the right, $n_{\max}$, where city size is so large with such enormous diseconomies that the population is indifferent between being in a rural settlement of size 1 (the size of a community formed by a defecting migrant) and $n_{\max}$ That is, given an inverted-U shape to real income I_1, self organization has cities at the right of the peak n^*, potentially at $n_{\max}$ where where $I_1\ (n=1) = I_1\ (n=n_{\max})$. The problem is the familiar one of co-ordination failure.

Consider a large economy with growing population, where, in size, all cities are at or just beyond n^*. Timely formation of the next city to accommodate this population growth requires en mass movement of population from existing cities into a new city of size, n^*. Without co-ordination in the form of developers or city governments, no such en mass movement is possible, so people wait to migrate

from existing cities to a new city until existing cities have all grown to $n_{\max}$, where it pays individual migrants to exit to a cities to set up their own tiny "city". At that "bifurcation point" (Krugman, 1991a), in equilibrium these milling migrants coalesce into 1 or more new cities of size greater than or equal to n^*, at which point, again, all then existing cities start to grow again with national population growth until they too hit the bifurcation point $n_{\max}$. This dismal process is what faces countries where local autonomy and national markets are poorly functioning, so that there are no market or institutional mechanisms to co-ordinate en mass movements of people. However the process we have outlined involving population swings across cities and potentially enormously over-populated cities may not be consistent with the data. In Section 3 we will outline a model with immobile capital, where self-organization can involve "commitment" given irreversibility of investment decisions. In that context outcomes, while still inefficient, are not so dismal.

2.1.2 Other City Types

In Black and Henderson, X_1 city type 1 is an input into production of the single final good in the economy, X_2 (from which, hence in a growth context human capital is also "produced"). In many models all outputs of specialized city types are final consumption goods. But here we follow Black and Henderson, without loss of generality. X_2 is produced in type 2 cities where the output for worker/firm j is correspondingly

$$X_{2j} = D_2 \ (n_2^{\delta_2} h_2^{\psi_2}) \ h_{2j}^{\theta_2} \ X_{1j}^{1-\alpha} \tag{16}$$

As in type 1 cities, per worker output is subject to own industry local scale externalities and to local knowledge spillovers. However now there is an intermediate input, X_{1j} , which is the numeraire good, with X_{2j} priced at P in national markets. The analysis of city sizes and formation for type 2 cities proceeds as for city type 1, with corresponding expressions, other than the addition of an expression for P in n_2^* and I_2 and a restriction for an inverted $U-$shape to I_2 that $\delta_2 < \alpha/2$.

In a static context the model is closed by utilizing the national full employment constraint

$$m_1 n_1 + m_2 n_2 = N \tag{17}$$

where m_1 and m_2 are the numbers of each type of city and N is national population. The second equation (to solve the 3 unknowns P, m_1 and m_2) equates real incomes as in equation (15) across cities

$(I_1 = I_2)$, where individual workers move across cities to equalize real incomes. Finally there is an equation where national demand equals supply in either the market. That is, the supply, $m_1 X_1$, equals the demand for X_1 as an intermediate input, $m_2 n_2 x_{1,}$ and for producing commuting costs $(m_1\ (bn_1^{3/2}) + m_2\ (bn_2^{3/2}))$ from eq. (9). In this specific model, the solution yields values of m_1, m_2 and *P* that are functions of parameters and h_1 and h_2. In a static context of identical workers, one would impose $h = h_1 = h_2$. We will discuss momentarily the solution for h_1 and h_2 and the model in the growth context. Later in section 3, we will detail solutions for prices and numbers of cities in a simpler but related two sector model. Here given log-linear production functions and a single final consumption good, as Black and Henderson show, X_1/X_2 and m_1 / m_2 will be constant over time, independent of *h.*

In the static context where, labor mobility requires $I_1 = I_2$, in the larger type of city, say type 1, commuting and land rent costs will be higher. Thus, if real incomes are equalized, from (15) , $W_1 > W_2$ as a compensating differential for higher living costs. Firms in type 1 cities are willing to pay higher wages because type 1 cities offer them greater scale benefits. Empirical evidence shows as cities move from a small size (say, 50,000) to very large metro areas, the cost-of-living typically doubles (Thomas 1978, Henderson, 2002), explaining the fact that nominal wages also double.

Another issue discussed at length in section 1.3 is that policy makers may favor large cities because they view them as "more productive". Indeed for an industry found in smaller towns, it may be that the externalities they face in equations (6) or (16) may be higher in a larger city. However that doesn't mean they locate there. Although externalities may be higher, in order for them to locate there, it must be sufficiently relatively higher to afford the higher wage and land rents, compared to a smaller city. If not, their profit maximizing or cost minimizing location is the smaller city.

Specialization. This analysis presumes cities specialize in production. That is an equilibrium outcome under a variety of conditions. In the model described so far, there are no costs of inter-city trade: no costs of shipping X_1 as inputs to X_2 type cities and shipping X_2 back as retail goods in X_1 type cities. All transport costs are internal to the city, given the relative greater importance of commuting costs in modern economies. Given that and given scale economies are internal to the industry, any specialized city (formed by a developer) out-competes any mixed city. The heuristic argument is simple. Consider any mixed city with $\tilde{n}_1$ and $\tilde{n}_2$ workers in industry 1 and 2. Split that city into two specialized cities, one with just $\tilde{n}_1$ people and the other with just $\tilde{n}_2$. Scale economies are undiminished ($\tilde{n}_1^{\delta_1}$ and $\tilde{n}_1^{\delta_2}$ in both cases in industries 1 and 2 respectively) but per worker commuting costs are lower in the specialized cities compared to the old larger mixed cities, so real incomes are higher in each specialized city compared to the old city.

Having own industry, or localization economies is a sufficient but not necessary condition for specialization. Industries can instead all have "urbanization" economies where scale depends on total local employment. However if the degree of urbanization economies differs across industries, then each industry has a different efficient local scale and is better off in a different size specialized city than any mixed city. Mixed cities occur more in situations where each good has localization economies enhanced by separate spillovers from the other industry or sharing of some common public infrastructure (Abdel-Rahman, 2000).

A basic problem in these urban models is the lack of nuance on transport costs. Either transport costs of goods across cities is zero or infinite as for housing, and potentially other non-tradables. A recent innovation is to have generalized transport costs (without a specific geography) where the cost of transporting a unit of X_1 to an X_2city is t_1 and the cost of shipping X_2 back to an X_1 city is t_2, an innovation due to Abdel-Rahman (1996) in a model similar to the one used here (one intermediate and one final good) and then generalized by Xiong (1998) and Anas and Xiong (1999). Now whether there are specialized as opposed to diversified cities depends on the level of t_1 and t_2. At appropriate points as t_1 or t_2 or both rise from zero, X_1 and X_2 will collocate (in developer run cities) in one type of city, while there may be some cities specialized in one of either X_1 or X_2. More generally with a spectrum of, say, final products, we would expect that some products with low enough t 's will always be produced in specialized cities, some high enough t 's will be in all cities, and some with middle range t 's will be produced in some cities (ones with bigger markets) but not others (with smaller markets). No one has yet simulated this more complex outcome.

2.1.3 Replicability and National Policy

At the national level in a large economy with many cities, at the limit, there are constant returns to scale, or replicability. If national population doubles, the numbers of cities of each type and national output of each good simply doubles, with individual city sizes, relative prices and real incomes unchanged .[5] With two goods and two factors, basic international trade theorems (Rybczynski, factor price equalization, and Stolper-Samuelson) hold (Hochman 1977, Henderson 1988). This gives an urban flavor to national policies (Renaud 1981, Henderson, 1988). For example trade protection policies favoring industry X_1 produced in relatively large size cities over industry X_2 produced in smaller type cities will alter national output composition towards X_1 production and increase the number of large relative to small cities. National urban concentration will rise. Similarly subsidizing an input such as

[5] Here with h_1 and h_2 yet to be solved we would need to double the numbers of people with h_1 and h_2 respectively. Below we will see the solution with growth to h_1 and h_2 is national scale invariant.

capital for a high tech product, X_1, again, say, produced in a larger type of city will cause the numbers of that type of city to increase, raising urban concentration.

2.2 Growth in a System of Cities

Black and Henderson (1999a) specify a dynastic growth model where dynastic families grow in numbers at rate g over time starting from size 1. If c is per person family consumption, the objective function is $\int_0^{\infty} \frac{c(t)^{1-\sigma}-1}{1-\sigma} e^{-(\rho-g)t} dt$ where $\rho\ (>g)$ is the discount rate. Dynasties can splinter (as long as they share their capital stock on an equal per capita basis) and the problem can be put in an overlapping generations context with equivalent results (Black, 2000), under a Galor and Zeira (1993) "joy of giving" bequest motive.

The only capital in the model is human capital and as such there is no market for it. Intra-family behavior substitutes for a capital market. Specifically families allocate their total stock of human capital (H) and members across cities, where Z proportion of family members go to type 1 cities (taking $Z\ h_1 e^{gt}$ of the H with them) and $(1-Z)$ go to type 2 cities taking $(1-Z)\ \mathrm{h}_2 e^{gt}$ with them). Additions to the family stock come from the equation of motion where the cost of additions, $P\dot{H}$, equals family income $Ze^{gt} I_1 + (1-Z)\ e^{gt} I_2)$ less the value of family consumption of X_2, or $Pc\ e^{gt}$. Constraints prohibiting consumption of human capital, non-transferability except to newborns, and non-transferability within families across city types (either directly or indirectly through migration) are non-binding on equilibrium paths.

Families allocate their populations across types of cities, with low human capital types (say h_1) "lending" some of their share $(h = H / e^{gt})$ to high human capital types (say h_2). High human capital types with higher incomes $(I_2 > I_1$ if $h_2 > h_1)$ repay low human capital types so $c_1 = c_2 = c$ (governed by the family matriarch). This in itself is an interesting development story, where rural families diversify migration destinations (including the own rural village) and remittances home are a substantial part of earnings. In Black and Henderson if capital markets operate perfectly for human capital (i.e., we violate the "no slavery" constraint) or capital is physical and capital markets operational, one dynastic family could move entirely to, say, type 1 cities and lend some of their human capital to another dynastic family in type 2 cities. With no capital market, each dynastic family must operate as its own informal capital market and spread itself across cities.

In this context Black and Henderson show that, regardless of scale or timing in the growth process, h_1 / h_2 and I_1 / I_2 are fixed ratios, dependent on θ_i in eq. (6) and (16). As θ_1 / θ_2 (the relative returns to capital) rises, h_1 / h_2, I_1 / I_2 and also n_1 / n_2 rise. Z and m_1 / m_2 are all fixed ratios of parameters

θ_i, δ_i, and α under equilibrium growth. Equilibrium and optimal growth differ because the private returns to education in a city, θ_i, differ from the social returns, $\theta_i + \psi_i$. But local governments can't intervene successfully to encourage optimal growth. Why? With free migration and "no slavery", if a city invests to increase its citizens' education, a person can take their human capital ("brain drain") and move to another city (be subsidized by another city to immigrate, given that city then need not provide extra education for that worker). This model hazard problem discourages internalization of education externalities.

2.2.1 Growth properties: Cities

From eq. (13), equilibrium (and efficient) city size in type 1 cities is a function of the per person human capital level, h_1, in type 1 cities. After solving out the model (for P), the same will be true of type 2 cities. City sizes grow as h_1 and h_2 grow, where, under equilibrium growth given h_1 / h_2 is a fixed ratio, $\dot{h}_1 / h_1 = \dot{h}_2 / h_2$ where a dot represents a time derivative. Then

$$\frac{\dot{n}_2}{n_2} = \frac{\dot{n}_1}{n_1} = 2\varepsilon_1 \frac{\dot{h}}{h} \tag{18}$$

where $\dot{n}_i / n_i$ is the growth rate of efficient sizes n_i^*.

For the number of cities, the issue is whether growth in individual sizes absorbs the national population growth, or more cities are needed. Given

$$\frac{\dot{m}_1}{m_1} = \frac{\dot{m}_2}{m_2} = g - \frac{\dot{n}_i}{n_i} = g - 2\varepsilon_1 \frac{\dot{h}}{h}, \tag{19}$$

the numbers of cities grow if $g > \dot{n}_i / n_i$. Note growth in numbers and sizes of cities is "parallel" by type, so the relative size distribution of cities is constant over time. Parallel growth with a constant relative size distribution of cities as reviewed in Section 1.1 is what is observed in the data. This result generalizes to many types of cities under certain conditions. For example, with the log linear production technologies we assumed and with many varieties of output consumed under unitary price and income elasticities of log-linear preferences, parallel growth results.

2.2.2 Growth properties: Economy

Ruling out explosive or divergent growth, there are two types of growth equilibria. Either the economy converges to a steady state level, or it experiences endogenous steady-state

growth. Convergence to a level occurs if $\varepsilon \equiv \varepsilon_1 \ (1-(\gamma - 2\delta_2)) + \varepsilon_2 \ (\gamma - 2\delta_2) < 1,$ where ε is a weighted average of the individual city type. In that case at the steady-state $\bar{h}$, $\dot{n}_i / n_i = 0$ and $\dot{m}_i / m_i = g$, or only the numbers but not sizes of cities grow just like in exogenous growth (Kanemoto 1980, Henderson and Ioannides, 1981). If $\varepsilon = 1$ then there is steady-state growth, where $\bar{\gamma}^h = \dot{h}/h = \frac{A-\rho}{\sigma}$ (where the transversality condition requires $A > \rho$). In that case $\dot{n}_i / n_i = 2\varepsilon_1^{\left(\frac{A-\rho}{\sigma}\right)}$, or cities grow at a constant rate. and their numbers also increase if $g > 2\varepsilon_1^{\left(\frac{A-\rho}{\sigma}\right)}$. This "knife-edge" formulation of whether there is endogenous growth or not dependent on the value of ε is not essential. For example in Rossi-Hansberg and Wright (2004) endogenous growth can occur more generally in a context where human capital accumulation involves worker time and the growth rate of human capital is a log-linear function of the fraction of time devoted to human capital accumulation, as opposed to production.

2.3 Extensions

There are three major extensions to the basic systems of cities models. First people may differ in terms of inherent productivity or in terms of endowments. Second, while we have discussed the issue of city specialization versus diversificationn, we haven't developed insights into a more nuanced role of small highly specialized cities versus large diversified metro areas in an economy.

2.3.1 Different Types of Workers

Turning to the first extension, Henderson (1974) has physical capital as a factor of production owned by capitalists who needn't reside in cities. Equilibrium city size reflects a market trade-off between the interests of city workers who have an inverted U-shape to utility as a function of the size of the city they live in and capitalists whose returns to capital rise indefinitely with city size (for the same capital to labor ratio). There is a political economy story, where capitalists collectively in an economy have an incentive to limit the number of cities, thus forcing larger city sizes. Helsley and Strange (1990) have a matching model between the attributes of entrepreneurs and workers and Henderson and Becker (2001) a related two class model. Again the two class model yields a conflict between the city sizes that maximize the welfare of one versus another group, which is resolved in competitive national land development markets.

In a different approach Abdel-Rahman and Wang (1997), Abdel-Rahman (2000) and later Black (2000) look at high and low skill workers who are used in differing proportions in production of different goods. Black has one traded good produced with just low skill labor and a second traded good produced with high skill workers and inputs of a local non-traded good produced with just low skill workers. High skill workers generate production externalities in the form of knowledge spillovers for all traded goods. In Black, urban specialization with all high skill workers (and some low skill workers) concentrated in one

type of city producing the first type of good is efficient; but a separating equilibrium that would sustain this pattern, where low skill workers and low tech production stay in their own type of city (rather than trying to cluster with high tech production) is not always sustainable. Black characterizes conditions under which a separating equilibrium will emerge.

It is important to note that there is a much more developed literature on inequality induced by neighborhood selection, where the characteristics of neighbors affect skill acquisition (e.g. average family background in the classroom affects individual student performance). That leads to segregation of talented or wealthier families by neighborhood (Benabou 1993, Durlauf, 1996) and can help transmit economic status across generations, promoting inter-generational income inequality.

2.3.2 Metro Areas.

Simple indices of urban diversity indicate that smaller cities are very specialized and larger cities highly diversified. So the question is what is the role of large metro areas in an economy and their relationship to smaller cities. Henderson (1988) and Duranton (2004) have a first nature - second nature world, where every city has a first nature economic base and footloose industries cluster in these different first nature cities. In general the largest centers are those attracting the most footloose production to their first nature center. The Duranton paper is discussed in more detail in Section 2.3.3. However, it seems that today few metro areas have an economic base of first nature activity. Accordingly recent literature has focused on the role of large metro areas as centers of innovation, headquarters, and business services (Kolko, 1999).

The Dixit-Stiglitz model opened up an avenue to look at large metro areas as having a base of diversified intermediate service inputs, which generate scale-diversity benefits for local final goods producers. That initial idea was developed in Abdel-Rahman and Fujita (1990) and has led to a set of papers focused on the general issue of what activities, under what circumstances are out-sourced. Theory and empirical evidence (Holmes, 1999 and Ono 2000) suggest that as local market scale increases, final producers will in-house less of their service functions. The resulting increased out-sourcing encourages competition and diversity in the local business service market, encouraging further out-sourcing.

In terms of incorporating this into the role of metro areas versus smaller cities, Davis (2000) has a two-region model, a coastal internationally exporting region and an interior natural resource rich region. There are specialized manufacturing activities which, for production and final sale, require business service activities, summarized as headquarters functions. Headquarters purchase local Dixit-Stiglitz intermediate services such as R&D, marketing, financing, exporting, and so on. Headquarters' activity is in port cities in the coastal region. The issue is whether manufacturing activities are also in these ports versus in specialized coastal hinterland cities versus in specialized interior cities. If the costs of interaction

(shipping manufactured goods to port and transactions costs of headquarters-production facility communication) between headquarters and manufacturing functions are extremely high, then both manufacturing and headquarters activities will be found together in coastal port cities. Otherwise they will be in separate types of cities where manufacturing cities will be in coastal hinterlands if costs of headquarters-manufacturing interaction are high, relative to shipping natural resources to the coast. However if natural resource shipping costs are relatively high, then manufacturing cities will be found in the interior. Duranton and Puga (2001) have a very similar model of functional specialization, without the regional flavor. If there is specialization, then there are headquarter cities where headquarters outsource local services in diversified large metro areas, while production occurs in specialized manufacturing cities.

In a different paper Duranton and Puga (2000) develop an entirely different and stimulating view of large metro areas. In an economy there are m types of workers who have skills each specific to producing one of m products. Specialized cities have one type of worker producing the standardized product for that type of worker subject to localization economies. Diversified cities have some of all types of workers. Existing firms at any instant die at an exogenously given rate; and, in a steady-state, new firms are their replacement. New firms don't know "their type" -- what types of workers they match best with and hence what final product they would be best off producing. To find their type they need to experiment by trying the different technologies (and hence trying different kinds of workers). New firms have a choice. They can locate in a diversified city with low localization economies in any one sector. But in a diversified city they can experiment with a new process each period until they find their ideal process. At that point they relocate to a city specialized in that product, with thus high localization economies for that product. Alternatively new firms can experiment by moving from specialized city to specialized city with high localization economies, but face a relocation cost each time. If relocation costs are high, it is best during their experimental period to be in a diversified city. This leads to an urban configuration of experimental diversified metro areas and other cities which are specialized in different standardized manufacturing products.

The Duranton and Puga model captures a key role of large diversified metro areas consistent with the data. They are incubators where new products are born and where new firms learn. Once firms have matured then they typically do relocate to more specialized cities. This also captures the product-life cycle for firms in terms of location patterns. Fujita and Ishii (1994) document the location patterns of Japanese and Korean electronics plants and headquarters. In a spatial hierarchy mega-cities house headquarters activities (out-sourcing business services) and experimental activity. Smaller Japanese or Korean towns have specialized, more standardized high tech production processes and low tech activity is off-shore.

2.3.3 Stochastic Process and Zipf's Law

Gabaix (1999a, 1999b) argues that if, there is a stochastic process where individual city growth rates follow Gibrat's Law—the growth rate in any period is unrelated to initial size -- then the size distribution that emerges will follow Zipf's Law. Beyond specifying a stochastic process where shocks to productivity or preferences follow a random walk, to get the result in a model where there is an endogenous number of cities of efficient sizes, as opposed to just fixing the number of cities (Gabaix 1999a, and Duranton 2004) requires considerable structure, with a variety of such issues being analyzed in Cordoba (2004). We follow Rossi-Hansberg and Wright (2004) who adapt the model we have presented. In their base case there is only human capital; and technology and preferences are log-linear. They have many final output industries and hence types of specialized cities. They group industries and specialized city types into sets. Within each set industries and city types have the *same* technology but each individual industry draws its own permanent shock each instant. In terms of the shock they assume that $D_1(t)$ in the equivalent of eq. (6) follows a finite order Markov process. Finally and critically to have Gibrat's Law lead to Zipf's Law, they must impose an arbitrary lower bound on the sizes that cities can fall to (Gabaix (1999a). These assumptions lead to Zipf's Law holding for each set of industries and they show one can aggregate across sets of industries to get Zipf's Law in aggregate. It goes without saying many of the assumptions imposed to get Zipf's Law are very strong, a key point made in Cordoba (2004).

In a recent paper, Duranton (2004) tries to model "micro-foundations" for the stochastic process affecting city sizes and as a result ends up modeling an important overlooked aspect of city evolution. Duranton has "first nature" (immobile given natural resource location) production and "second nature" (mobile, or footloose) production in m cities, where m is given by the number of immobile natural resource products, each needing their own city. So, in contrast to Rossi-Hansberg and Wright the number of cities is fixed; but given that restriction a lot is accomplished. In the paper there are $(n >> m)$ products, in a Grossman-Helpman (1991) product quality ladder model. The latest innovation in each product is produced by the monopolist holding the patent and only this top quality is marketed for any product. Investment in innovation to try to move the next step up in the quality ladder in industry k and get the next patent in k, can also lead to the next step up in a different industry -- i.e., there can be cross-industry innovation. For footloose industries, to partake of a winning innovation occurring for industry k in city i, requires industry k production to locate in city i where the innovator is. Presumably co-location of the inventor and production makes the information needed for the transition to mass production cheaper to exchange (e.g., the workers in the innovative firm take over production). Innovation follows a stochastic process where innovation probabilities depend on R&D expenditures. Industry jumps from city to city according to where the latest innovation is, and city growth also follow a stochastic process. The resulting stochastic process of city growth and decline results in steady state size distributions that are similar to

Zipf's Law. Adding in considerations of urban scale economies in the innovation process helps explain the long right tails in actual city size distributions, as they differ from Zipf's Law.

Duranton's formulation has the nice feature that cities have patterns of production specialization which change over time. This seems to fit the data; and Duranton's paper in fact models the evolution of industry structure of cities. We know from Black and Henderson (1999b) and Ellison and Glaeser (1999) that industries move "rapidly" across cities, with city specialization changing over time for cities. Any city is very slow to gain a high share of any particular industry's production (given there are many possible industries to gain a share from) and is very quick to lose a high share (given many competitor cities).

3. Urbanization and Growth

The previous section examined a fully urbanized economy where all production occurs in cities. City sizes grow with improvements in technology; but, absent stochastic elements, individual cities grow in parallel, with the relative numbers of different types of cities and the relative size distribution of cities time invariant. Here we examine a non-steady state world in which an economy has an agriculture sector that is shrinking with economic development and an urban sector that is growing. We briefly review traditional dual sector models and the new economic geography models, both of which examine sectoral transformation, but without cities per se and generally without economic growth. Then we present an endogenous growth model in which there is sectoral change with cities.

3.1 Two Sector Approaches, Without Cities

Urbanization involves resources shifting from an agricultural to an urban sector. The dual economy models dating back to Lewis (1954) look at sectoral change but are really static models. They focus on the question of urban " bias", or the effect of government policies on the urban-rural divide, and the efficient rural-urban allocation of population at a point in time. These two sector models have an exogenously given "sophisticated" urban sector and a "backward" rural sector (Rannis and Fei 1961, Harris and Todaro, 1970, and others as now well exposited in textbooks (e.g., Ray 1998)).

In these models, the marginal product of labor in the urban sector is assumed to exceed that in the rural sector. Arbitrage in terms of labor migration is limited by inefficient (and exogenously given) labor allocation rules such as farm workers being paid average rather than marginal product or artificially limited absorption in the urban sector (e.g., formal sector minimum wages). The literature focuses on the effect on migration from the rural to urban sector of policies such as rural-urban terms of trade, migration restrictions, wage subsidies, and the like.

The final and most complex versions of dual sector models are in Kelley and Williamson (1984) and Becker, Mills, and Williamson (1992), which are fully dynamic CGE models. They have savings behavior and capital accumulation, population growth, and multiple economic sectors in the urban and rural regions. Labor markets within sector and across regions are allowed to clear. The models analyze the effects of a wider array of policy instruments, including sector specific trade or capital market policies for housing, industry, services and the like. However the starting point is again an exogenously given initial urban-rural productivity gap, sustained initially by migration costs and exogenous skill acquisition. On-going urbanization is the result of exogenous forces -- technological change favoring the urban sector or changes in the terms of trade favoring the urban sector.

As models of urbanization, these dual economy ones are a critical step but they suffer obvious defects. First how the dual starting point arises is never modeled. Second, and related to the first, there are no forces for agglomeration that would naturally foster industrial concentration in the urban sector. Finally although the models have two sectors there is really little spatial or regional aspect to the problem. There is a new generation of two-sector models, the core-periphery models, which attempt to address some of these defects. The core-periphery models ask under what conditions in a two-region country, industrialization, or "urbanization" is spread over both regions versus concentrated in just one region.

Compared to the dual economy models, Krugman's (1991a) paper explicitly has scale economies that foster endogenous regional concentration. Second, while there are two regions, no starting point is imposed, where one region is assumed to start off ahead of the other. Industrialization may occur in both regions or in only one region. One region can become "backward" (under certain assumptions), or, if not backward (lower real incomes) at least relatively depopulated (Puga, 1999). But these are outcomes solved for in the model. Third the models have some notion of space represented as transport costs of goods between regions.

The models are focused on a key developmental issue -- the initial development of a core (say, coastal) region and a periphery (say, hinterland) region, as technology improves (transport costs fall) from a situation starting with two identical regions. As such they do relate to the earlier discussion in Section 1.3 of urban concentration in a primate city versus the rest of the urban sector. Some work (Puga 1999, Fujita, Krugman, and Venables 1999, Chapter 7, Helpman 1998, and Tabuchi 1998) also analyzes how under certain conditions, with further technological improvements, there can be reversal. Some industrial resources leave the core; and the periphery also industrializes/urbanizes. However core-periphery models have limited implications for urbanization per se, since in many versions including Krugman's (1991a) initial paper, the agricultural population is fixed.

Unfortunately, to date core-periphery models have been almost exclusively uni-dimensional in focus, asking what happens to core-periphery development as transport costs between regions decline.

They are not focused on other forms of technological advance, let alone endogenous technological development. With a few exceptions such as Fujita and Thisse (2002) and Baldwin (2001), the models are static. But even in these exceptions, there is still the focus on exogenous changes in transport technology. Compared to the older dual economy literature, generally core-periphery models have no policy considerations of interest to development economists, such as the impact of wage subsidies, rural-urban terms of trade, capital market imperfections. An exception is that some papers have examined the impact on core-periphery structures of reducing barriers to international trade, such as tariff reduction; and papers are starting to explore issues of capital market imperfections. The core-periphery model is an important innovation in bringing back the role of transport costs, largely ignored in urban systems work, to the forefront. Excellent summaries of the key elements include Neary (2001), Fujita and Thisse (2000) and Ottaviano and Thisse (2004), with the latter two developing many extensions. Fujita, Krugman and Venables (1999) stands as a basic reference on detailed modeling.

The dual economy and core-periphery models are regional models, with limited urban implications. Urban models are focused on the city formation process, where the urban sector is composed of numerous cities, endogenous in number and size. Efficient urbanization and growth require timely formation of cities. As policy issues the extent of market completeness in the national markets in which cities form, the role of city governments and developers, the role of inter-city competition, and the role of debt finance and taxation are critical. In the next section we analyze an urbanization process in which there are cities. Then we turn to a discussion of some key policy issues.

3.2 Urbanization With Cities

Here I present a simple two sector model of urbanization with cities, adapting the model in part 2 following Henderson and Wang (2005, 2004). The urban sector is exactly like the X_1 city sector earlier, with production technology given in (6). The other sector is food produced in the agriculture sector, which we make now the numeraire (since there may initially be no urban sector). As a result, for type 1 cities in the urban sector, equation (7) for wages, equations (9) and (10) for commuting costs and rents, and equation (15) for income are all redefined to be multiplied by the price of X_1, p. The city size equation is the same, invariant to relative prices. Critical here is

$$I_1 = W_1\ (1-2\delta_1) = pQ_1 h_1^{\varepsilon_1} \qquad (15a)$$

for Q_1 a parameter cluster.

Agriculture. Rural output per worker is $D_a h_a^{\psi_a} h_a^{\theta_a}$, or $D_a h_a^{\ \varepsilon_a}$, so rural wages and real income are

$$W_a = D_a h_a^{\varepsilon_a} \qquad \varepsilon_a \equiv \psi_a + \theta_a \tag{20}$$

As such the rural sector is very simple: no commuting costs, no agglomeration economies and no diminishing returns to land. As in the urban sector, productivity is affected by individual human capital accumulation, $h_a^{\theta_a}$, and by sector knowledge spillovers, $h_a^{\psi_a}$.

Preferences and Urbanization. To have sectoral transformation we need to move away from the world of unitary price and income elasticities in Section 2, so growth between sectors is not parallel. Here we assume preferences have the form

$$V = (x + a^{\gamma})^{\alpha}, \qquad \gamma, a < 1 \tag{21}$$

where *a* is consumption of agricultural products. In (21) agricultural demand is income inelastic, with a demand function

$$a = \gamma^{\frac{1}{1-\gamma}} p^{\frac{1}{1-\gamma}}. \tag{22}$$

3.2.1 Human Capital Market, Migration, Savings.

The urbanization process as a "transitory" phenomenon is not a steady state process. To simplify, following much of the literature, we introduce an explicit market for human capital, as though human capital investments were not embodied. And we assume an exogenous savings rate, s. For the former, now each person in the economy has a human capital level, h, which can be used in production or can be loaned out. I now flesh out the equations of the model, that in Section 2 were skimmed over. The capital market equalizes capital returns across sectors so $r = p\ \theta_1\ D_1\ h_1^{\theta_1+\psi_1-1}\ n_1^{\delta_1} = \theta_a\ D_a\ h_a^{\theta_a+\psi_a-1}$. Substituting in for n_1 from (13)

$$p = Q_2\ h_a^{\varepsilon_a} h_1^{1-\varepsilon_1} \tag{23}$$

recalling $\varepsilon_a \equiv \psi_a + \theta_a$ and $\varepsilon_1 \equiv (\theta_1 + \psi_1)/(1 - 2\delta_1)$. Q_2 is a parameter cluster.

Free migration requires net urban incomes including capital costs to equal the same for agriculture, or $I_1 + r\ (h - h_1) = W_a + r\ (h - h_a)$. Utilizing (15)

$$W_1 - W_a = pbn^{1/2} + r\ (h_1 - h_a) \tag{24}$$

With free migration equalizing real incomes across sectors, urban wages exceed rural wages by (commuting) cost-of-living differences (the first term on the RHS of (24)), and by a factor compensating if human capital requirements in the urban sector exceed those in the rural, as I assume.

If we substitute in (24) for $W_1,\ W_a,\ p$ and r and rearrange, we get

$$h_a = h_1 \ \frac{\theta_a}{\theta_1} \left(\frac{1-\theta_1-2\delta}{1-\theta_a} \right) \tag{25}$$

A sufficient condition for $h_1 > h > h_a$, or the urban sector to be human capital intensive, is that $\theta_1 > \theta_a$, as assumed.

To close the model requires three relationships. First is national full employment of capital and labor so

$$n_a h_a + n_1 m_1 h_1 = hN \tag{26a}$$

$$n_a + m_1 n_1 = N \tag{26b}$$

where m_1 ,as before, is the number of type 1 cities and N is the national population. The third equation equates the demand for food equal to its supply. But that requires a digression on how human capital is produced and nature of savings. Since we want to be able to start with a purely rural economy, we don't want to have it produced just from X_1 as in Section 2.

We assume human capital production in each sector is made from goods from that sector (where an equal expenditure in any sector results in the same human capital), which is almost like assuming, for a fixed savings rate, a fixed fraction of working time in any sector is needed to produce a unit of human capital. Second, we assume savings at the rate s are from wage income net of rental costs, or from $I_1 - rh_1,$ and $W_a - rh_a$, which magnitudes are equalized by migration. Thus in the food market total production $n_a D_a h_a^{\varepsilon_a}$ equals food consumption demand $N\ (p\gamma)^{\frac{1}{1-\gamma}}$ (see 22) plus agricultural savings, or $n_a D_a h_a^{\varepsilon_a} = N\ (p\gamma)^{\frac{1}{1-\gamma}} + sn_a\ (W_a - rh_a)$. Substituting in for $r,$ for p from (22), W_a from (20) and for h_a from (25) we get

$$n_a / N = Q_3 h_1^{\frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma}} \tag{27}$$

with Q_3 a parameter cluster. We assume $\gamma\varepsilon_a - \varepsilon_1 < 0$, so the social returns to human capital in the urban sector exceed those in the rural sector discounted by γ. With economic growth in human capital, the rural sector diminishes. Note in (27) for there to be an urban sector, h_1 must be large enough so $n_a / N < 1$, as we explain below. Of course h_1 is linked to h through (26a) where with substitutions

$$h_1 \left(1 - Q_4 \, h_1^{\frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma}} \right) = h \tag{28}$$

where given $\gamma\varepsilon_a - \varepsilon_1 < 0$, $dh_1 / dh > 0$, once there is an urban sector. Q_4 is a parameter cluster.

3.2.2 Urban Growth and Transformation

Once an urban sector exists, city growth is as in section 2: $\dot{n}_1 / n_1 = 2\varepsilon_1 \dot{h}_1 / h_1$, so cities grow with human capital accumulation. The growth in number of cities now depends on the rate of urbanization, as well. Combining (26a) and (26b), with differentiation $\dot{m}_1 / m_1 = (N / m_1 n_1)\, g - \dot{n}_1 / n_1 - (n_a / m_1 n_1)\, \dot{n}_a / n_a$. If we differentiate (27) for $\dot{n}_a / n_a$ and combine this becomes

$$\dot{m}_1 / m_1 = g - \dot{n}_1 / n_1 - \frac{n_a}{m_1 n_1} \, \frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma} \, \dot{h}_1 / h_1 \tag{29}$$

As before, the rate of growth of numbers of cities is increased by national population growth, g, and reduced by growth in individual city sizes. Now it is also enhanced by economic growth which increases relative demand for urban products and draws labor out of agriculture, as captured by the last term in (29).

3.2.3 Economic Growth

Given the savings rule, total human capital increases by $\dot{H} = s\left[m_1 n_1 \, (I_1 - r h_1) + n_a \, (W_a - r h_a) \right]$ each instant so the per person change in capital is $\dot{h} / h = \dot{H} / H - g$. Given $I_1 - r h_1 = W_a - r h_a$, with substitutions we have

$$\dot{h} / h = s Q_5 h_1^{\varepsilon_a - 1} \left(1\text{-}Q_6 h_1^{\frac{\gamma\varepsilon_a - \varepsilon_1}{1-\gamma}} \right)^{-1} - g \tag{30}$$

where $Q_6 < Q_4$, for parameter clusters. In terms of growth, if $\varepsilon_a < 1$ and urbanization occurs, we have steady state levels given $\dot{h}/h$ declines with increases in h_1, and hence *h*. If $\varepsilon_a = 1$, we approach steady state growth once h_1 gets large so $n_a N \to 0$ and the expression in parentheses in (30) approaches 1. However in either case at low levels of development in (27), n_a / N is bounded at 1 where equation (27) defines a critical h_1 and hence h in (28), say h_c, below which $n_a / N = 1$. To have steady state levels with urbanization given we start at $n_a / N = 1$ with $\dot{h}/h = \dot{h}_a / h_a = -g + (1-\theta_a)\, h_a^{\varepsilon_a - 1}$, requires $h_c < (g/(1-\theta_a))^{\frac{1}{\varepsilon_a - 1}}$, so we pass the critical h at which urbanization starts before hitting the potential steady state value of *h* without urbanization. Otherwise the economy can be stuck with no urbanization. Details of this and issues of multiple equilibria are discussed in Henderson and Wang (2005).

3.3 Extensions and Policy Issues

There are three general sets of policy issues. First concerns whether in the context of the models in section 2 and 3.2, the national composition of cities of different types is efficient. We have already discussed this issue: in many contexts asking whether the national composition of cities is efficient is the same as asking if national output composition is efficient. If there are national policy biases such as trade policies favoring steel products over textile products, with urban specialization, if steel is produced in bigger types of cities than textiles, the numbers of larger cities relative to smaller ones and hence urban concentration will increase. The second set of policy issues concerns whether, in general, city sizes are likely to be efficient and we discuss this in Section 3-3-1.

The second general set of issues deals with factors we have ignored. In particular the modeling in sections 2 and 3.2 assumes a nice smooth process where (i) all factors of production are perfectly mobile and malleable, (ii) city borrowing and debt accumulation have no role, (iii) "lumpiness" problems that arise in city formation when economies are small are ignored: while m must be an integer in reality, in the analysis it is treated as any positive number where the number of cities grows at a rate $\dot{m}/m$, rather than by 0, 1 or 2. A model that incorporates these features is outlined in section 3.3.2, which brings to the forefront a variety of policy issues.

3.3.1 City Sizes

A perpetual debate in particular developing countries is whether certain mega-cities are oversized, squandering national resources that must be allocated to commuting, congestion, and transport in those cities and resulting in low quality of life in the polluted, unsanitary and crowded slums of such cities. In other countries, especially former planned economies, the debate goes the other way: are cities too small? The growth connection is straightforward. Either squandered resources in over-sized cities or too small cities with unexploited scale economies mean lower income levels, potentially lower savings,

lower capital accumulation and thus lower growth rates. While calculations are tedious, in the steady state growth in section 2.2.2 where $\gamma^h = \dot{h}/h = (A-\rho)/\sigma$, A depends on urban parameters (for example, increasing with human capital returns) and will be lowered if city sizes are inefficient.

Using a simple, partial equilibrium diagram, it is possible to illustrate both issues: the mega-city "problem" and the planned economy problem. The diagrams point to first order effects. For the mega-city problem, suppose there are a variety of type 1 cities in an economy with free mobility of labor and institutions supporting efficient city formation. In Figure 5a, the representative city has a size n_1^*, where real income as function of city size peaks at I_1^*, tangent to the perfectly elastic national supply curve of labor to the city, given perfect labor mobility. Suppose one particular type 1 city is favored relative to the rest, where various types of favoritism are discussed in Section 1.3.2. For example it may have special public services compared to other cities financed out of national taxes. Those favors raise the realized utility, or real income that residents in the favored city potentially receive, shifting up the inverted-U real income curve. That upward shift draws migrants into the city expanding its size to n_{mega}. But at n_{mega}, the net income *generated by* the city, ignoring its nationally financed favors, is only net I_m. The gap, $I_1^* - I_m$, times the population represents "squandered resources". Of course, such squandering would in general equilibrium affect prices, lowering the height of the population supply curve and the inverted-U's.

A second issue in city formation concerns poor institutions in national land markets and in local governance which limit the number of cities that can form. Suppose that, in villages which might become cities, local governments by institutional restrictions can't expand infrastructure (see next section), can't rezone and build on urban fringe land, and can't offer subsidies to incoming firms. And suppose developers can't assemble large tracts of land for development because property rights are ill-defined. These villages can't grow into cities; as well, entirely new cities can't form. If the number of cities is bindingly limited, so there are too few cities, all existing cities under free migration are too big. In Figure 5a, suppose we reconsider the figure ignoring the representative city curve and assume all cities have inverted-U's like the favored city. Then, in this reinterpretation of the figure, I_F is the potentially attainable real income in all cities (ignoring general equilibrium effects) if cities could freely form. Given restricted numbers, rather than operating at I_F (with size n_1^*), cities are overcrowded; and in equilibrium they operate at , say, I_1^* (with size n_{mega}), with the same national supply curve of labor as labeled in the figure. The restrictions result in losses related to the gap $I_F - I_1^*$.

The planned economy problem is entirely different. Former "planned" economies like China have formal migration restrictions limiting the visas given for rural people to move to cities and limiting migrants' access to jobs, housing, medical care and schooling in destination cities to reduce the incentive

to migrate. Some former planned economies (as well as China) limited migration through housing provision and land development. If the state provides and allocates all housing assignments, migrants can't move unless housing is provided in the destination. As we saw in Table 2, countries like China and Russia have very low urban concentration compared to other large countries. Figure 5b captures the essence of the problem. While the representative city has an inverted-U where real income is maximized at n_1^*, migration restrictions for cities a and b restrict sizes to n_a and n_b and real incomes to I_a and I_b. Au and Henderson (2004) estimate these inverted-U's for different types of cities in China in 1997 and find that 30% of cities are significantly undersized – below the lower 95% confidence interval on their equivalent to n_1^*. The productivity losses from being undersized are enormous: 30-50% or more loss in GDP per capita for many cities.

3.3.2 Sequential City Formation and Governance

In a working paper, Henderson and Venables (2004) take a new approach to city formation. They assume a context where (1) there is a steady-flow of migrants from rural to urban areas, and (2) urban residence requires a fixed investment in non-malleable, immobile capital (housing, sewers, water mains, etc.). Cities form sequentially without population swings, so migrants all flow first into city 1 until its equilibrium size is reached (abstracting from any on-going technological change), and then all future migrants all go to a second city until its equilibrium size is reached, and so on. This is a very different process than when all resources are mobile. In the usual models in a small economy, when the second city forms, it takes half the population of the first at that instant, and when the third forms it takes one third of the then population of the first two. Cities grow way past $\overset{*}{n}_1$, shrink back to $\overset{*}{n}_1$, and then grow again, shrink, and on so. With fixed capital, such population swings would mean periods of abandoned housing. With sufficiently high required fixed capital investments, all population swings are eliminated in equilibrium. Each new city starts off tiny with no accumulated scale effects and low productivity. It grows steadily absorbing all new rural-urban migrants until its growth interval is complete and it reaches steady state size; then a new city starts off growing from a tiny size.

With sequential city formation without population swings, given discounting of the future, efficient city size requires cities to grow past the equivalent of n_1^* to their steady state size, n_{opt}, at which point real income per worker is declining. Intuitively, growing past n_1^*, with declining but still high real income, postpones the formation of a new city with tiny population, no scale effects and very low incomes. The paper then looks at equilibrium city formation in two contexts.

First is a situation with no "large agents" in national land markets – no developers and no city governments. In a model with perfect mobility of resources as discussed in Section 2, city formation with atomistic agents is a disaster due to coordination failure. A new city can only form when old cities are so

big that the income levels they offer have fallen to the point where they equal what a person can earn in a city of size one. Having immobile capital presents a commitment device (Helsley and Strange, 1996), so individual, sequentially rational builders switch from building in an old city to building in a new one at a "reasonable time". Real incomes are still equalized across cities through migration. Given big old cities have high nominal incomes and the tiny new one low nominal income, housing rents adjust in old cities to equalize real incomes. Housing rents in old cities change over the growth cycle of a new city, starting very high and then declining (see also Glaeser and Gyourko, 2003). In this context, equilibrium city sizes may even be smaller than optimal ones. The deviation from optimum has not to do with coordination failure which is solved despite the absence of "large" agents, but with the present value of externalities created by the marginal migrant in an old versus a new city.

With developers or full empowered local governments, externalities are appropriately internalized and city sizes are n_{opt}. However, apart from financing the housing and infrastructure capital, to induce new migrants to move to a new city with its low real income and scale economies in a timely fashion, large agents must subsidize in-migration of worker-firms. To do this they must borrow and, in fact, public debt accumulates over the entire growth interval of a city and only starts to be paid off once it reaches steady state size. Debt ceilings, or limits for cities which are common in many countries curtail subsidies to in-migrants and postpone new city formation. Debt limited cities are too big. The paper also explores the effects of limits on local tax property tax powers.

4. Some Issues For a Research Agenda

A handbook paper is a place to offer research suggestions, as well as summarize the state of knowledge. While various avenues of research are noted throughout, here I summarize three key suggestions. In all the spatial and urban work, transport costs are either absent or treated as a technology parameter that may exogenously change. In an actual development and growth context, transport costs reflect public infrastructure investment decisions, subject to political influence. Models need to endogenize transport costs, so spatial structures across regions and cities are an outcome of investment decisions. A similar comment involves mobility costs of workers, which are related to both transport and communication infrastructure investments.

A second key research issue involves spatial inequality as it evolves with growth in a context where workers have different ability endowments and choose different human capital levels. In most spatial and urban models, workers are identical, except for their degree of mobility. But with

urbanization, it may be that it is higher ability rural folks who urbanize and acquire modern skills, increasing real income gaps between high and low ability people. We have no models that directly address these issues and provide a comprehensive framework to evaluate spatial inequality, or cross-space income differences.

Finally we don't really have models that address the evolution of city production patterns with on-going technological change. While we have looked at parallel growth and urbanization, we know city functions also change over time. In less developed countries, bigger cities may be focused on manufacturing, but somehow with growth and technological change, big cities tend to specialize more in service functions, purchased by manufacturers and retailers in smaller cities. While we have models of functional specialization , we haven't modeled this evolution in city roles over the development process.

References

Abdel-Rahman, H. (1996), "When do Cities Specialize in Production," Regional Science and Urban Economics, 26 1-22.

Abdel-Rahman, H. (2000), "City Systems: General Equilibrium Approaches," in J-M Huriot and J-F Thisse (eds.), Economics of Cities: Theoretical Perspectives, Cambridge University Press, 109-37.

Abdel-Rahman, H. and A. Anas (2004), "Theories of Systems of Cities," in J.V. Henderson and J-F. Thisse (eds.), Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

Abdel-Rahman, H. and M. Fujita (1990), "Product Variety, Marshallian Externalities, and City Sizes," Journal of Regional Science, 30, 165-85.

Abdel-Rahman, H. and P. Wang (1997), "Social Welfare and Income Inequality in a System of Cities," Journal of Urban Economics, 41, 462-83.

Ades, A.F. and E.L. Glaeser (1995), "Trade and Circuses: Explaining Urban Giants," Quarterly Journal of Economics, 110, 195-227.

Anas, A. and K. Xiong, (1999), "The Formation and Growth of Specialized Cities," State University of New York at Buffalo mimeo.

Arthur (1990), "'Silicon Valley Locational Clusters: When Do Increasing Returns to Scale Imply Monopoly," Mathematical Social Sciences, 19, 235-51.

Au, C.C. and J.V. Henderson (2004), "How Migration Restrictions Limit Agglomeration and Productivity in China," Brown University mimeo.

Baldwin, R.E. (2001), "Core-Periphery Model with Forward-Looking Expectations," Regional Science and Urban Economics, 31, 21-49.

Barro, R. and X. Sala-i-Martin (1991), "Convergence Across States and Regions," Brookings Papers on Economic Activity, 1,107-82.

Barro R. and X. Sala-i-Martin (1992), "Regional Growth and Migration: A Japan-USA Comparison," Journal of Japanese and International Economics, 6, 312-46.

Becker, G., E. Mills, J.G. Williamson (1992), Indian Urbanization and Economic Growth Since 1960, Johns Hopkins Press.

Beeson, P.E., D.N. DeJong, and W. Troeskan (2001), "Population Growth in US Counties, 1840-1990," Regional Science and Urban Economics, 31, 669-700.

Benabou, R. (1993), "Workings of a City: Location, Education, and Production," Quarterly Journal of Economics, 108, 619-52.

Bergsman, J., P. Greenston, and R. Healy (1972), "The Agglomeration Process in Urban

Growth," Urban Studies, 9, 263-88

Black, D. (2000), "Local Knowledge Spillovers and Inequality," University of California Irvine mimeo.

Black, D. and J.V. Henderson (1999a), "A Theory of Urban Growth, " Journal of Political Economy, 107, 252-84.

Black, D. and J.V. Henderson (1999b), "Spatial Evolution of Population and Industry in the USA," Papers and Proceedings of the American Economic Association, May.

Black, D. and J.V. Henderson (2003), "Urban Evolution in the USA," Journal of Economic Geography, 3,343-373.

Clark, J.S. and J.C. Stabler (1991), "Gibrat's Law and the Growth of Canadian Cities," Urban Studies, 28, 635-39.

Cordoba, J-C (2004), "On the Size Distribution of Cities," Rice University mimeo.

Davis, James (2000), "Headquarter Service and Factory Urban Specialization With Transport Costs," Brown University mimeo.

Davis, J. and J.V. Henderson (2003), "Evidence on the Political Economy of the Urbanization Process," Journal of Urban Economics, 53, 98-125.

Dixit A. and J. Stiglitz (1977), "Monopolistic Competition and Optimum Product Diversity," American Economic Review, 67, 297-308.

Dobkins, L.H. and Y.M. Ioannides (2001), "Spatial Interactions Among U.S. Cities: 1900-1990," Regional Science and Urban Economics, 31, 701-32.

Duranton, G. (2004), "City Size Distribution as a Consequence of the Growth Process," LSE mimeo.

Duranton, G. and H. Overman (2004), "Testing for localization using micro-geographic data," Revised working paper, April 2004.

Duranton, G. and D. Puga (2000), "Nursery Cities: Urban Diversity Process Innovation, and the Life Cycle of Products," American Economic Review, 91, 1454-77.

Duranton, G. and D. Puga (2001), "From Sectoral to Functional Urban Specialization", CEPR, LSE Discussion Paper 2971.

Duranton, G. and D. Puga (2004), "Microfoundations of Urban Agglomeration Economies" in Handbook of Urban and Regional Economics, Volume 4, J.V. Henderson and J-F Thisse (eds.), North Holland.

Durlauf, S.N. (1996), "A Theory of Persistent Income Inequality," Journal of Economic Growth, 1, 75-93.

Eaton, J. and Z. Eckstein (1997), "Cities and Growth: Evidence from France and Japan,"

Regional Science and Urban Economics, 27, 443-74.

Ellison, G. and E. Glaeser (1999a), "The Geographic Concentration of US Manufacturing: A Dartboard Approach," Journal of Political Economy, 105, 889-927.

Ellison, G. and E. Glaeser (1999b), "The Geographic Concentration of Industry: Does Natural Advantage Explain Agglomeration," American Economic Association Papers and Proceedings, 89, 311-16.

Fay, M. and C. Opal (1999), "Urbanization Without Growth: Understanding an African Phenomenon," World Bank mimeo.

Flatters, F., J.V. Henderson and P. Mieszkowski (1974), "Public Goods, Efficiency, and Regional Fiscal Equalization," Journal of Public Economics, 3, 99-112.

Fujita, M. and H. Ogawa (1982), "Multiple Equilibria and Structural Transition of Non-Monocentric Configurations," Regional Science and Urban Economics, 12, 161-96.

Fujita, J., P. Krugman and A.J. Venables (1999), The Spatial Economy: Cities, Regions, and International Trade, MIT Press.

Fujita, M. and T. Ishii (1994), "Global Location Behavior and Organization Dynamics of Japanese Electronic Firms and Their Impact on Regional Economies," Paper presented for Prince Bertil Symposium on the Dynanmic Firm, Stockholm.

Fujita, M. and J-F Thisse (2000), "The Formation of Economic Agglomerations," in J-M Huriot and J-F Thisse (eds.) Economies of Cities, NY, Cambridge University Press.

Fujita, M. and J-F Thisse (2002), Economics of Agglomeration, Cambridge University Press.

Gabaix, X. (1999a), "Zipf's Law and the Growth of Cities," American Economic Association and Proceedings, 89, 129-32.

Gabaix, X.(1999b), "Zipf's Law for Cities: an Explanation," Quarterly Journal of Economics, 114, 739-67.

Gallup, J.L., J.D. Sacks and A. Mellinger (1999), "Geography and Economic Development," International Regional Science Review, 22, 179-232.

Galor, O. and J. Zeira (1993), "Income Distribution and Macro Economics," Review of Economic Studies, 60, 35-52.

Glaeser, E., J. Scheinkman and A. Schelifer (1995), "Economic Growth in a Cross-Section of Cities," Journal of Monetary Economics, 36, 117-34.

Glaeser, E. and J. Gyourko (2003), “Urban Decline and Durable Housing,” Harvard University mimeo.

Grossman, G. and E. Helpman (1991), "Quality Ladders in the Theory of Growth," Review of Economic Studies, 58, 43-61.

Harris, J. and M. Todaro (1970), "Migration, Unemployment and Development: A Two Sector Analysis," American Economic Review, 40, 126-42.

Head, K and T. Mayer (2004), "The Empirics of Agglomeration and Trade*," in J.V. Henderson and J-F Thisse (eds.), Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

E. Helpman (1998), "The Size of Regions," in D. Pines, E. Sadka and I. Zilcha (eds.), Topics in Public Economics: Theoretical and Applied Analysis, Cambridge University Press, 33-54.

Helsley, R. and W. Strange (1990), "Matching and Agglomeration Economies in a System of Cities," Regional Science and Urban Economics, 20, 189-212.

Helsley, R. and W. Strange (1994), "City Formation with Commitment," Regional Science and Urban Economics, 24, 373-390.

Henderson, J.V. (1974), "The Sizes and Types of Cities," American Economic Review, 61, 640-56.

Henderson, J.V. (1988), Urban Development: Theory, Fact and Illusion, Oxford University Press.

Henderson, J.V. (1997), "Medium Size Cities", Regional Science and Urban Economics, 27, 449-470.

Henderson, J.V. (2002), "Urban Primacy, External Costs, and Quality of Life," Resource Economics and Energy, 24, 95-106.

Henderson, J.V. (2003) "The Urbanization Process and Economic Growth: The So-What Question," Journal of Economic Growth, 8, 47-71.

Henderson, J.V. and Y. Ioannides (1981), "Aspects of Growth in a System of Cities," Journal of Urban Economics, 10, 117-39.

Henderson, J.V. and A. Kuncoro (1996), "Industrial Centralization in Indonesia," World Bank Economic Review 10, 513-40.

Henderson, J.V., A. Kuncoro and P. Nasution (1996), "Dynamic Development in Jabotabek," Indonesian Bulletin of Economic Studies, 32, 71-96.

Henderson, J.V. and R. Becker (2001), "Political Economy of City Sizes and Formation," Journal of Urban Economics, 48, 453-84.

Henderson, J.V., T. Lee and J.Y. Lee (2001), "Scale Externalities in Korea," Journal of Urban Economics, 49, 479-504.

Henderson, J.V. and A.J. Venables (2004), "The Dynamics of City Formation: Finance and Governance," LSE mimeo.

Henderson, J.V. and H.G. Wang (2004), "Urbanization and City Growth," Brown University mimeo.

Henderson, J.V. and H.G. Wang (2005), “Urbanization and City Growth,” Journal of Economic Geography, forthcoming

Hochman, O. (1977), "A Two Factor Three Sector Model of an Economy With Cities," mimeo.

Holmes, T. (1999), "Localization of Industry and Vertical Disintegration," Review of Economics and Statistics, 81, 314-25.

Ioannides, Y.M. and H.G. Overman (2003), “Zipf's Law for Cities: An Empirical Examination," Regional Science and Urban Economics, 33, 1, March, 127-137.

Junius, K. (1999), "Primacy and Economic Development: Bell Shaped or Parallel Growth of Cities," Journal of Economic Development, 24 (1), 1-22.

Kanemoto, Y. (1980), Theories of Urban Externalities, Amsterdam: North-Holland.

Kelly, A.C. and J.G. Williamson (1984), What Drives Third World City Growth? A Dynamic General Equilibrium Approach, Princeton University Press.

Kim, H.S. (1988), "Optimal and Equilibrium Land Use Pattern in a City: A Non-Parametric Approach," Ph.D. thesis, Brown University.

Kim, S. (1995), "Expansion of Markets and the Geographic Distribution of Economic Activities: The Trends in US Manufacturing Structure, 1860-1987," Quarterly Journal of Economics, 95, 881-908.

Kolko, J. (1999), "Can I Get Some Service Here? Information Technology Service Industries, and the Future of Cities," Harvard University mimeo.

Krugman, P. (1991a), "Increasing Returns and Economic Geography," Journal of Political Economy, 99, 483-99.

Krugman, P. (1991b), Geography and Trade, MIT Press, Cambridge.

Lee, K.S. (1988), "Infrastructure Constraints on Industrial Growth in Thailand," World Bank INURD Working Paper No. 88-2.

Lee, K.S. (1989), The Location of Jobs in a Developing Metropolis, Oxford University Press.

Lee, T.C. (1997), "Industry Decentralization and Regional Specialization in Korean Manufacturing," unpublished Brown University Ph.D. thesis.

Lewis, W.A. (1954), "Economic Development With Unlimited Supplies of Labor," Manchester School of Economic and Social Studies, 22, 139-91.

Lucas, R.E. (1988), "On the Mechanics of Economic Development," Journal of Monetary Economics, 12, 3-42.

Lucas, R.E. and E. Rossi-Hansberg (2002), "On the Internal Structure of Cities," Econometrica, 70:4, 1445-1476.

Marshall, A. (1890), Principles of Economics, London: MacMillan.

Mills, E. and C. Becker (1986), Studies in Indian Urban Development, Oxford University Press.

Mills, E. and B. Hamilton (1994), Urban Economics, Scott-Foresman.

Mohring, H. (1961), "Land Values and Measurement of Highway Benefits," Journal of Political Economy, 49, 236-49.

Neary, J.F. (2001), "Of Hype and Hyperbolas: Introducing the New Economic Geography," Journal of Economic Literature, 49, 536-61.

Ono, Y. (2000), "Outsourcing Business Service and the Scope of Local Markets," CES Discussion Paper CES 00-14.

Ottaviano, G, and J-F Thisse (2004), "Agglomeration and Economic Geography," in J.V. Henderson and J-F Thisse (eds) Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

Overman, H., S. Redding and A. J. Venables (2003), "The Economic Geography of Trade, Production and Income: A Survey of Empirics," Handbook of International Trade, J. Harrigan and K. Choi, (eds.), Blackwell.

Puga, D. (1999), "The Rise and Fall of Regional Inequalities," European Economic Review, 43, 303-34.

Quah, D. (1993), "Empirical Cross Section Dynamics and Economic Growth," European Economic Review, 37, 426-34.

Rannis, G. and J. Fei (1961), "A Theory of Economic Development," American Economic Review, 51, 533-65.

Rappaport, J. and D. Sacks (2003), "The US as a Coastal Nation," Journal of Economic Growth, 8, 5-46.

Rauch, J.E. (1993), "Does History Matter Only When It Matters a Little? The Case of City-Industry Location," Quarterly Journal of Economics, 108, 843-67.

Ray, D. (1998), Development Economics, Princeton: Princeton University Press.

Renaud, B. (1981), National Urbanization Policy in Developing Countries, Oxford University Press.

Rosen, K. and M. Resnick (1980), "The Size Distribution of Cities: An Examination of the Pareto Law and Primacy," Journal of Urban Economics, 81, 165-86.

Rosenthal, S. and W. Strange (2004), "Evidence on the Nature and Sources of Agglomeration Economies" in J.V. Henderson and J-F Thisse (eds.) Handbook of Urban and Regional Economics, Vol. 4, Cities and Geography, North Holland.

Rossi-Hansberg, E. (2004), "Optimal Urban Land Use and Zoning," Review of Economic Dynamics,7, 69-106.

Rossi-Hansberg, E. and E.M. Wright (2004), “Urban Structure and Growth”, Stanford University (May) mimeo

Simon, H. (1995), "On a class of skew distribution functions," Biometrika, 44, 425-40.

Stiglitz, J. (1977), "The Theory of Local Public Goods," in M.S. Feldstein and R.P. Inman (eds.), The Economics of Public Services, London: MacMillan, 273-334.

Tabuchi, T. (1998), "Urban Agglomeration and Dispersion: A Synthesis of Alonso and Krugman," Journal of Urban Economics, 44, 333-351.

Thomas, V. (1978), “The Measurement of Spatial Differences in Poverty: The Case of Peru,” World Bank Staff Working Paper No. 273.

Wheaton, W. and H. Shishido (1981), "Urban Concentration, Agglomeration Economies, and the Level of Economic Development," Economic Development and Cultural Change, 30, 17-30.

Williamson, J. (1965), "Regional Inequality and the Process of National Development," Economic Development and Cultural Change, June, 3-45.

World Bank (2000), Entering the 21st Century: World Development Report 1999/2000, Oxford University Press.

Xiong, K. (1998), "Intercity and Intracity Externalities in a System of Cities: Equilibrium, Transient Dynamics, and Welfare Analysis," unpublished Ph.D. thesis, State University of New York at Buffalo.

Table 1. World City Size Distribution, 2000

size range	**count**	**mean**	**Share[1]**
17,000,000 < =n2000	4	20,099,000	4.5
12,000,000 < =n2000 < 17,000,000	7	13,412,714	5.2
8,000,000 < =n2000 < 12,000,000	13	10,446,385	7.5
4,000,000 < =n2000 < 8,000,000	29	5,514,207	8.9
3,000,000 < =n2000 < 4,000,000	41	3,442,461	7.8
2,000,000 < =n2000 < 3,000,000	75	2,429,450	10.1
1,000,000 < =n2000 < 2,000,000	247	1,372,582	18.8
500,000 < =n2000 < 1,000,000	355	703,095	13.9
250,000 < =n2000 < 500,000	646	349,745	12.5
100,000 < =n2000 < 250,000	1,240	157,205	10.8
Overall	2,657	658,218	100.0

1) a ratio of total population in the group to total population of cities with > =100,000

Table 2. Spatial Inequality

	1960		**2000**		
	(1)	**(2)**	**(3)**	**(4)**	**(5)**
	Gini	**Number of Cities**	**Gini**	**Number of Cities**	**Rank Size Coefficient "a"**
World	.59	1197	.56	1673	n.a.
Developed	.65	523	.58	480	n.a.
Soviet bloc	.52	179	.45	202	n.a.
Less developed	.57	495	.56	991	n.a.
Brazil	.67	24	.65	65	-.87
China	.47	108	.43	223	-1.3
India	.56	95	.58	138	-1.1
Indonesia	.52	22	.61	30	-.90
Mexico	.61	28	.60	55	-1.04
Nigeria	.31	20	.60	38	-.98
France	.61	31	.59	27	.97
Germany	.6	44	.56	31	-.74
Japan	.60	100	.66	82	-1.06
Russia	.54	79	.46	91	-1.34
Spain	.53	27	.52	20	-.98
Ukraine	.44	25	.40	32	-1.31
UK	.68	39	.60	21	-.83
USA	.58	167	.54	197	-1.11

Table 3. Total Numbers of Cities and Sizes

	1960	**1970**	**1980**	**1990**	**2000**
number of cities	969	1,129	1,353	1,547	1,568
mean size	556,503	640,874	699,642	789,348	943,693
median size	252,539	275,749	304,414	355,660	423,282
minimum size	100,082	115,181	126,074	141,896	169,682

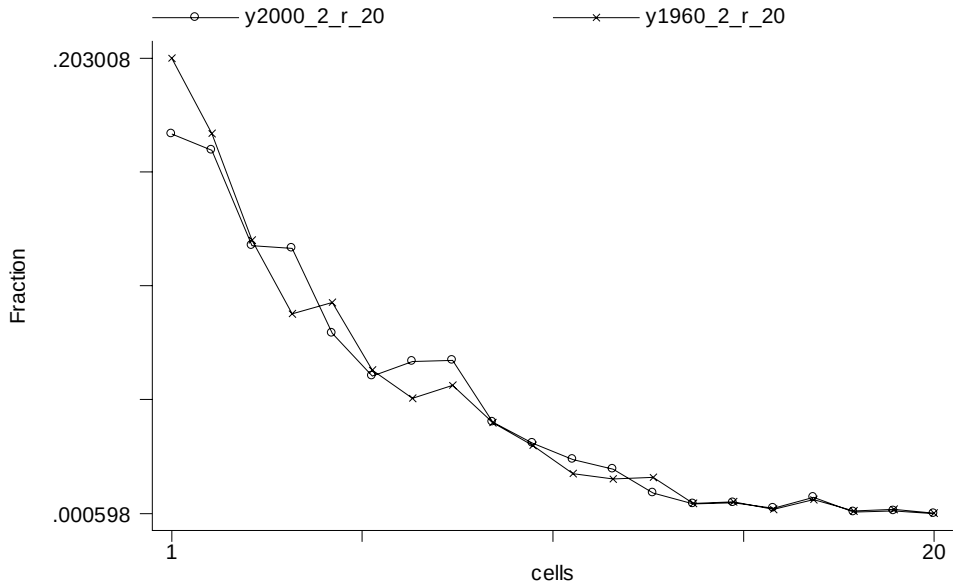


Figure 1a. Relative Size Distribution of Cities for all Countries

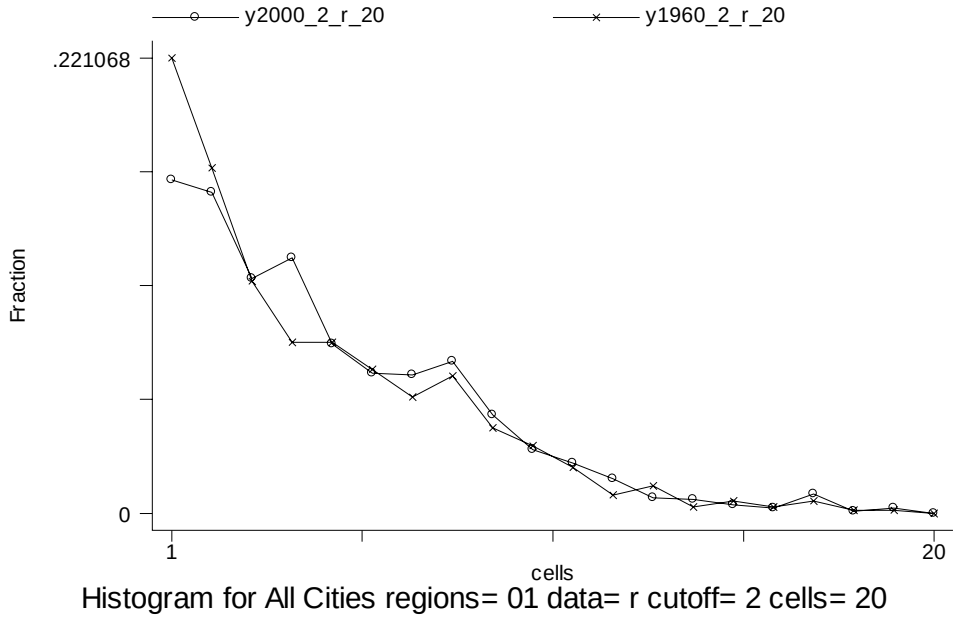


Figure 1b. Relative Size Distribution of Cities in Developing and Transition Economies

Figure 2: Share of Urban Population in Total Population.
(average over countries within groups)

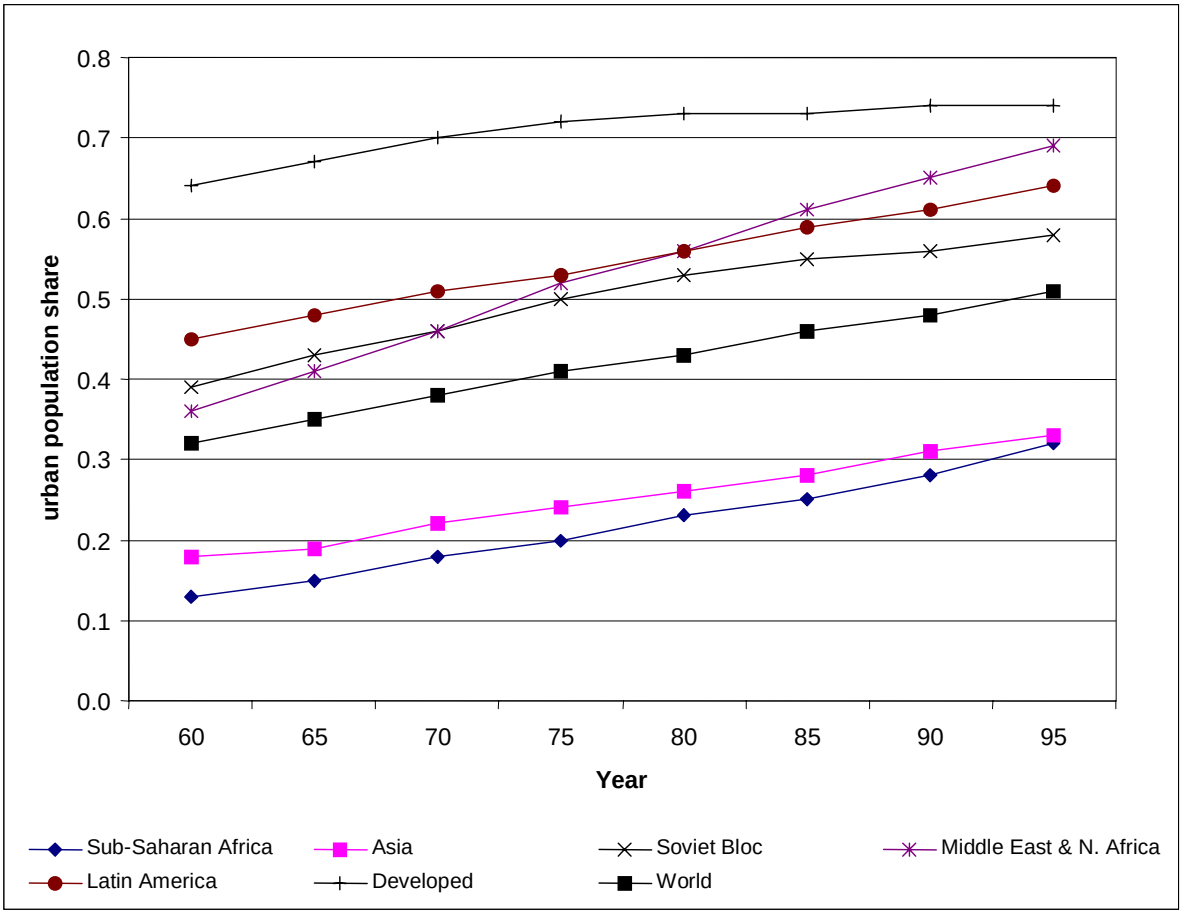


Figure 3: Partial Correlation Between Ln(urban population) and Ln(real GDP per capita), Controlling for Ln(national population), 1965-1995.

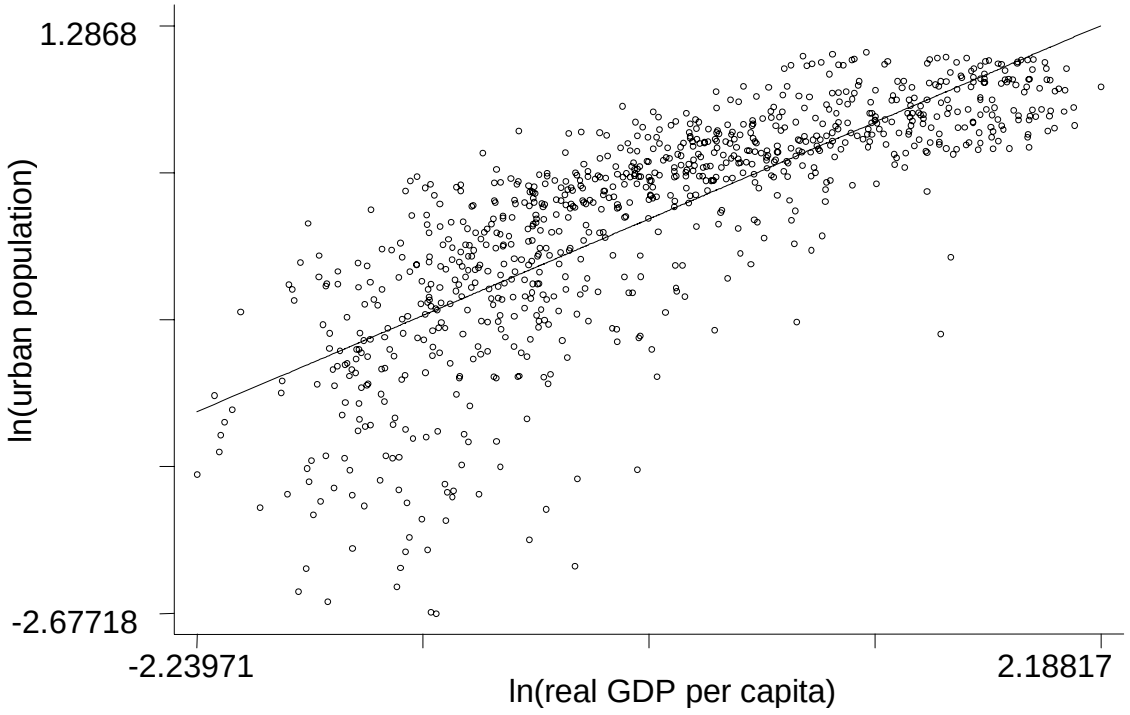
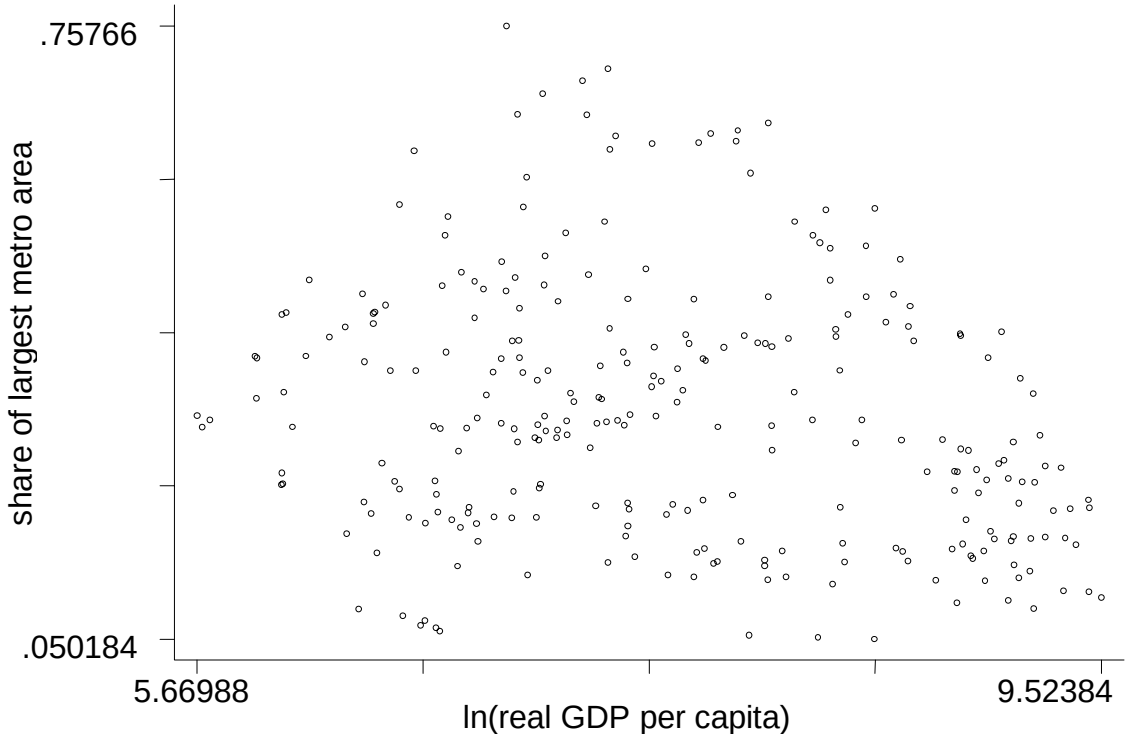


Figure 4: Primacy and Economic Development.

(a) Early period: 1965-75.



(b) Recent Period: 1985-95.

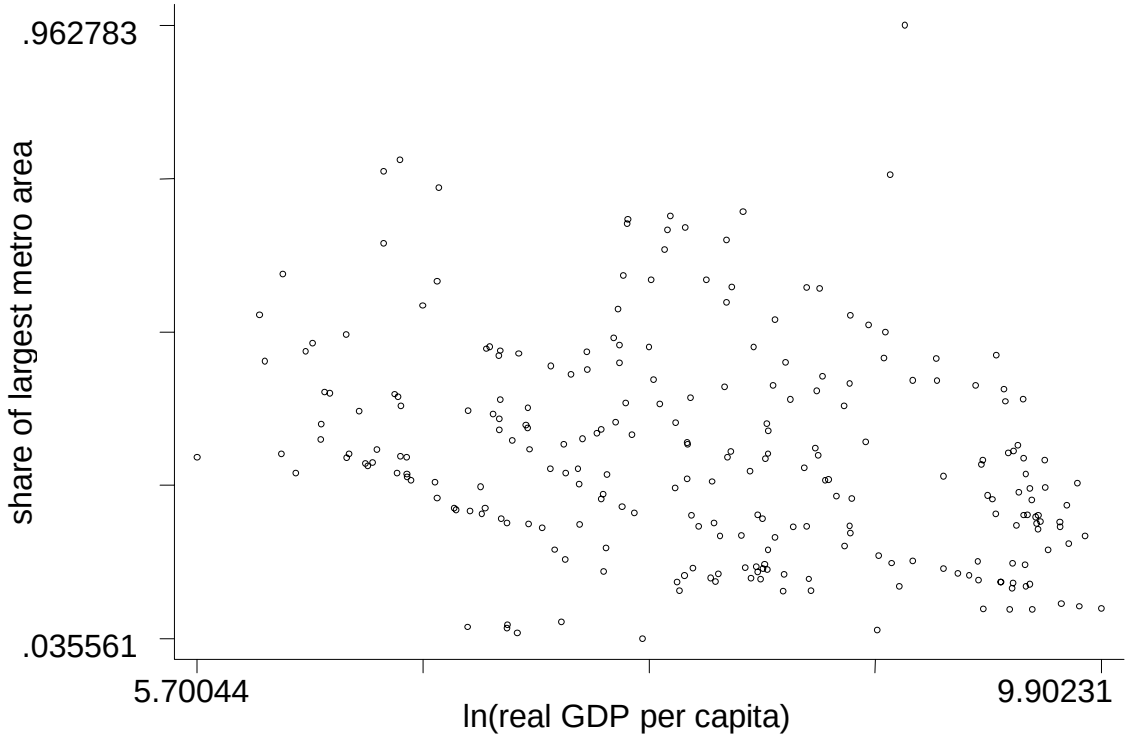
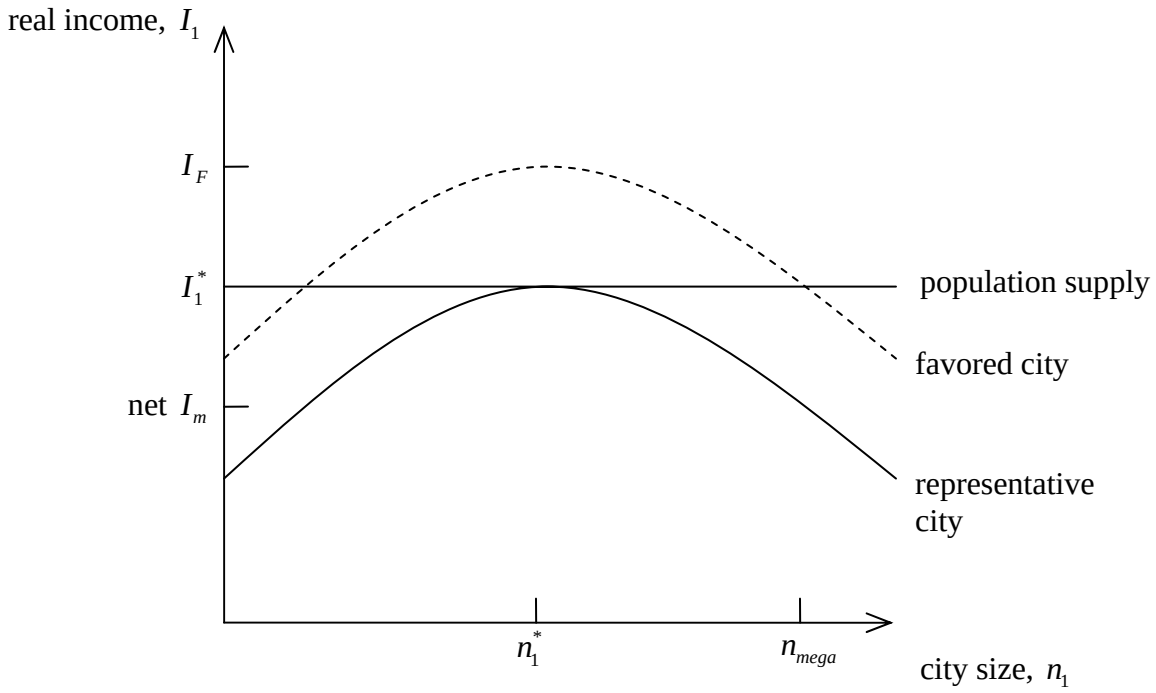
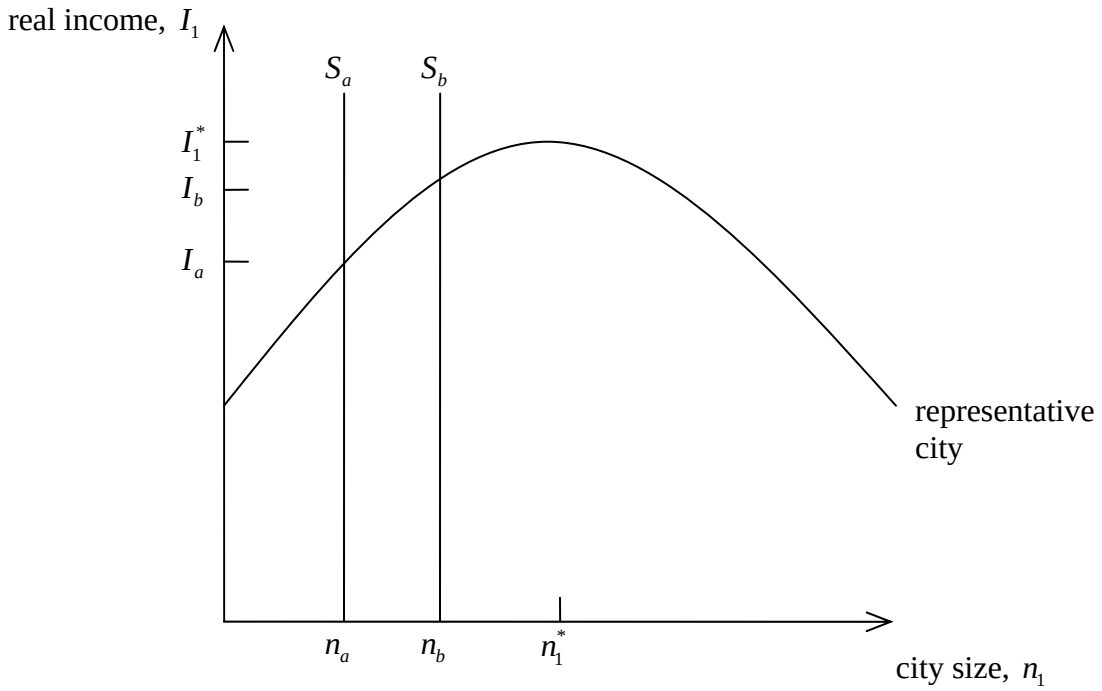


Figure 5. City Sizes



(a) Favored Cities



(b) Migration Restrictions